John Searle

Mentes, cerebros y ciencia

Traducción de Luis Valdés

CATEDRA
TEOREMA

Título original de la obra: *Minds, Brains and Science*
The 1984 Reith Lectures

© Jonh R. Searle
Ediciones Cátedra, S. A., 1985
Don Ramón de la Cruz, 67. 28001-Madrid
Depósito legal: M. 43.801-1985
ISBN: 84-376-0569-5
Printed in Spain
Impreso en Lavel
Los Llanos, nave 6. Humanes (Madrid)

Índice

A Dagmar

INTRODUCCIÓN

Fue un gran honor para mí el que se me pidiera dar las *Reith Lectures* de 1984. Desde que Bertrand Russell comenzó la serie en 1948, éstas son las primeras que da un filósofo.

Pero si dar las conferencias es un honor, es también un desafío. La serie ideal de las conferencias Reith debe consistir en seis unidades radiofónicas, cada una de exactamente media hora de duración, cada una con entidad independiente de modo que pueda ser autosuficiente, aunque cada una contribuya al todo unificado que consta de las seis. La serie debe construirse sobre el trabajo previo del conferenciante, pero al mismo tiempo debe contener material nuevo y original. Y quizá lo más difícil de conseguir, debe ser completamente asequible a un auditorio interesado y atento, muchos de cuyos miembros no tienen familiaridad alguna con el tema, con su terminología o con las especiales preocupaciones de sus profesionales. No sé si todos estos objetivos son simultáneamente alcanzables, pero en cualquier caso son aquellos a los que aspiraba. Una de mis razones de mayor peso para querer dar las conferencias Reith fue la convicción de que los resultados y los métodos de la filosofía analítica moderna pueden hacerse asequibles a un auditorio mucho más amplio.

Mis primeros planes para la versión en forma de libro consistían en ampliar cada uno de los capítulos, de modo que intentaran hacer frente a todas las objecciones que yo pudiera imaginar que vendrían de mis quisqui-

llosos compañeros filósofos, por no mencionar a los colegas de la ciencia cognitiva, inteligencia artificial y otros campos. Mi plan original era, dicho brevemente, intentar convertir las conferencias en un libro convencional con notas a pie de página y todo lo demás. Al final tomé la decisión de no hacer esto precisamente porque ello destruiría lo que para mí era, en primer lugar, una de las cosas más atractivas de la serie: su completa accesibilidad a cualquiera que estuviese suficientemente interesado en intentar seguir las argumentaciones. Estos capítulos son, pues, las conferencias Reith tal como las pronuncié. He ampliado alguna en aras de una mayor claridad, pero he intentado mantener el estilo, el tono y la informalidad de las conferencias originales.

El tema predominante de esta serie son las relaciones de los seres humanos con el resto del universo. Específicamente, la cuestión de cómo reconciliamos una cierta tradición mentalista que tenemos de nosotros mismos con una concepción aparentemente inconsecuente del universo como un sistema puramente físico, o un conjunto de sistemas físicos en interacción. Alrededor de este tema cada capítulo está dirigido a una cuestión específica: ¿cuál es la relación de la mente con el cerebro? ¿Pueden los computadores digitales tener mentes en virtud solamente de que tienen los programas correctos con los *inputs* y los *outputs* correctos? ¿Hasta qué punto es plausible el modelo de la mente como un programa de computador? ¿Cuál es la naturaleza de la estructura de la acción humana? ¿Cuál es el estatus de las ciencias sociales en tanto que ciencias? ¿Cómo podemos reconciliar, si es que podemos, nuestra convicción de nuestro libre albedrío con nuestra concepción del universo como un sistema físico, o un conjunto de sistemas físicos en interacción?

Durante la elaboración de la serie surgieron otros ciertos temas importantes que no pudieron ser completamente desarrollados, simplemente por las limitaciones del formato. Quiero hacerlos completamente explícitos en esta introducción, y, haciéndolo así, pienso que puedo ayudar al lector a comprender mejor los capítulos que siguen.

El primer tema es cuán poco sabemos del funcionamiento del cerebro humano y cuánto dependen de esta ignorancia las pretensiones de ciertas teorías. Como David Hubel, el neurofisiólogo escribió en 1978: «Nuestro conocimiento del cerebro está en un estado muy primitivo. Mientras que para algunas regiones hemos desarrollado algún género de concepto funcional, hay otras, del tamaño de un puño, de las que como mínimo puede decirse que estamos en el mismo estado de conocimiento que estábamos respecto al corazón antes de que nos diésemos cuenta de que bombeaba sangre.» Y ciertamente, si el lego interesado selecciona media docena de libros de texto estándar sobre el cerebro, como yo hice, y los aborda en un esfuerzo para obtener respuestas al tipo de cuestiones que se le ocurrirían inmediatamente a cualquier persona curiosa, probablemente se sentiría decepcionado. ¿Qué es exactamente la neurofisiología de la conciencia? ¿Por qué necesitamos dormir? ¿Por qué exactamente nos emborracha el alcohol? ¿Cómo exactamente se almacenan los recuerdos en el cerebro? Cuando escribo esto, simplemente no sabemos las respuestas a ninguna de esas cuestiones fundamentales. Muchas de las afirmaciones sobre la mente hechas en varias disciplinas que van desde la psicología freudiana a la inteligencia artificial dependen de este tipo de ignorancia. Tales afirmaciones viven en los agujeros de nuestro conocimiento.

Según la explicación tradicional del cerebro, la explicación que considera a la neurona como la unidad fundamental del funcionamiento del cerebro, la cosa más destacable de dicho funcionamiento es simplemente ésta. Toda la enorme variedad de *inputs* que recibe el cerebro —los fotones que golpean la retina, las ondas sonoras que estimulan el tímpano, el estímulo sobre la piel que activa las terminaciones nerviosas para la presión, el calor, el frío, el dolor, etc.—, todos esos *inputs* se convierten en un medio común: proporcionan variables de activación de las neuronas. Además, e igualmente destacable, esas proporciones variables de activación de las neuronas en diferentes circuitos neuronales y diferentes condiciones locales del cerebro, producen toda

la variedad de nuestra vida mental. El olor de una rosa, la experiencia del azul del cielo, el sabor de la cebolla, el pensamiento de una fórmula matemática: todo esto es producido por proporciones variables de activación de las neuronas, en diferentes circuitos, con relación a diferentes condiciones locales del cerebro.

Ahora bien, ¿cuáles son exactamente esos diferentes circuitos neuronales y cuáles son los diferentes ambientes locales que dan cuenta de las diferencias en nuestra vida mental? Nadie lo sabe con detalle, pero tenemos pruebas de que determinadas regiones del cerebro están especializadas en ciertos tipos de experiencias. El córtex visual juega un papel especial en las experiencias visuales, el córtex auditivo en las experiencias auditivas, etcétera. Supongamos que el córtex visual fuese alimentado con estímulos auditivos y el córtex auditivo lo fuese a su vez con estímulos visuales. ¿Qué sucedería? Hasta donde yo sé nadie ha hecho jamás el experimento, pero parece razonable suponer que el estímulo auditivo sería 'visto'; esto es, produciría experiencias visuales, y el estímulo visual sería 'oído'; esto es, produciría experiencias auditivas, y ambas cosas sucederían a causa de las características específicas, aunque ampliamente desconocidas, del córtex visual y auditivo, respectivamente. Aunque esta hipótesis es especulativa, tiene algún apoyo independiente si se reflexiona sobre el hecho de que un puñetazo en el ojo produce un destello visual ('ver las estrellas'), aunque no es un estímulo óptico.

Un segundo tema que recorre desde el principio hasta el fin todos estos capítulos es que tenemos una resistencia cultural heredada a tratar la mente consciente como un fenómeno biológico igual que cualquier otro. Esto se remonta a Descartes en el siglo XVII. Descartes dividió el mundo en dos géneros de substancias: substancias mentales y substancias físicas. Las substancias físicas eran el dominio propio de la ciencia y las substancias mentales eran propiedad de la religión. Algo parecido a una aceptación de esta división existe incluso en la actualidad. Así, por ejemplo, la conciencia y la subjetividad se consideran a menudo como tópicos inadecuados para la ciencia. Y esta reluctancia a tratar de

la conciencia y de la subjetividad es parte de una persistente tendencia objetivante. La gente piensa que la ciencia debe tratar sobre fenómenos objetivamente observables. En ocasiones en que he dado conferencias a auditorios de biólogos y neurofisiólogos, he encontrado que muchos de ellos eran muy reacios a tratar la mente en general y la conciencia en particular como un dominio propio de la investigación científica.

Un tercer tema que recorre, subliminalmente, todos estos capítulos es que la terminología tradicional que tenemos para discutir esos problemas es inadecuada en varios sentidos. De los tres términos que componen el título *Mentes, cerebros y ciencia,* solamente el segundo está completamente bien definido. Por 'mente' entiendo solamente las secuencias de pensamientos, sensaciones y experiencias, conscientes e inconscientes, que componen nuestra vida mental. Pero el uso del nombre 'mente' está habitado peligrosamente por los fantasmas de viejas teorías filosóficas. Es muy difícil resistirse a la idea de que la mente es un género de cosa, o al menos una especie de plaza de toros, o al menos algún género de caja negra, en la que ocurren todos esos procesos mentales.

La situación de la palabra 'ciencia' es incluso peor. De buena gana prescindiría de esta palabra si pudiese. 'Ciencia' se ha convertido en algo parecido a un término honorífico, y todo tipo de disciplinas que son completamente distintas de la física y de la química están ansiosas de llamarse a sí mismas 'ciencias'. Una buena regla práctica a tener presente es que cualquier cosa que se llame a sí misma 'ciencia' probablemente no lo es —por ejemplo, la ciencia cristiana, o la ciencia militar y probablemente incluso la ciencia cognitiva o la ciencia social—. La palabra 'ciencia' tiende a sugerir un numeroso grupo de investigadores en bata blanca agitando tubos de ensayo y mirando de cerca los instrumentos. Para muchas mentes esto sugiere una infalibilidad arcana. El cuadro rival que quiero sugerir es éste: a lo que aspiramos en las disciplinas intelectuales es al conocimiento y a la comprensión. Ya lo tengamos en matemáticas, crítica literaria, historia, física o filosofía, hay solamente conocimiento y comprensión. Algunas disci-

plinas son más sistemáticas que otras, y podríamos querer reservar la palabra 'ciencia' para ellas.

Estoy en deuda con un número más bien amplio de estudiantes, colegas y amigos por su ayuda en la preparación de las *Reith Lectures* tanto por lo que respecta a la versión radiada como a ésta en forma de libro. Quiero dar las gracias especialmente a Alan Code, Rejane Carrion, Stephen Davies, Hubert Dreyfus, Walter Freeman, Barbara Horan, Paul Kube, Karl Pribam, Gunther Stent y Vanessa Whang.

La B.B.C. fue excepcionalmente comprensiva. George Fisher, el director del *Talks Departament*, fue un apoyo extraordinario, y mi productor, Geoff Deehan, fue simplemente excelente. Mis mayores deudas son con mi esposa, Dagmar Searle, que me ayudó en todos y cada uno de los pasos del camino y a quien está dedicado este libro.

1

EL PROBLEMA MENTE-CUERPO

Durante miles de años la gente ha estado intentando comprender sus relaciones con el resto del universo. Por diversas razones muchos filósofos son reacios hoy en día a abordar tan grandes problemas. Sin embargo, los problemas permanecen y en este libro voy a hacer frente a algunos de ellos.

Por el momento, el mayor problema es éste: tenemos una cierta representación de sentido común de nosotros mismos como seres humanos que es muy difícil casar con nuestra concepción 'científica' global del mundo físico. Nos pensamos a nosotros mismos como agentes *conscientes, libres, cuidadosos, racionales* en un mundo del que la ciencia nos dice que consta enteramente de partículas físicas carentes de mente y de significado. Ahora bien. ¿Cómo podemos conjugar esas dos concepciones? ¿Cómo, por ejemplo, puede ser el caso de que el mundo no contenga otra cosa que partículas físicas inconscientes y que, con todo, contenga también conciencia? ¿Cómo puede un universo mecánico contener seres humanos intencionales —esto es, seres humanos que pueden representarse el mundo a sí mismos? ¿Cómo, para decirlo brevemente, puede un mundo esencialmente carente de significado contener significados?

Tales problemas se vierten sobre otras cuestiones que suenan más contemporáneas: ¿Cómo debemos interpretar el trabajo reciente en informática e inteligencia arti-

ficial, trabajo que aspira crear máquinas inteligentes? Específicamente, ¿nos da el computador digital la representación correcta de la mente humana? ¿Y por qué sucede que las ciencias sociales en general no nos han permitido formarnos ideas sobre nosotros mismos comparables a las ideas que las ciencias naturales nos han permitido formarnos sobre el resto de la naturaleza? ¿Cuál es la relación entre las explicaciones ordinarias, de sentido común, que aceptamos, sobre los modos en que la gente se comporta y los modos científicos de explicación?

En este primer capítulo quiero zambullirme directamente en lo que muchos filósofos piensan que es el problema más difícil de todos: ¿Cuál es la relación de nuestras mentes con el resto del universo? Este, estoy seguro que se reconocerá, es el problema tradicional mente-cuerpo o mente-cerebro. En su versión contemporánea toma usualmente la forma: ¿cómo se relaciona la mente con el cerebro?

Creo que el problema mente-cuerpo tiene una solución más bien simple, una solución que es coherente tanto con lo que sabemos de neurofisiología, como con nuestra concepción de sentido común acerca de la naturaleza de los estados mentales : dolores, creencias, deseos y así sucesivamente. Pero antes de presentar esa solución, quiero preguntar por qué el problema mente-cuerpo parece tan intratable. ¿Por qué tenemos todavía en filosofía y en psicología, después de todos esos siglos, un 'problema mente-cuerpo' en un sentido en que no tenemos, por así decirlo, un 'problema digestión-estómago'? ¿Por qué parece la mente más misteriosa que otros fenómenos biológicos?

Estoy convencido de que parte de la dificultad es que nos empeñamos en hablar sobre un problema del siglo XX en un vocabulario anticuado del siglo XVII. Cuando yo era estudiante de los primeros cursos de carrera, recuerdo que estaba insatisfecho con las elecciones de las que aparentemente se disponía en filosofía de la mente: se podía ser o monista o dualista. Si se era monista se podía ser o materialista o idealista; si se era materialista se podía ser o conductista o fiscalista. Y así sucesivamente. Una de mis aspiraciones en lo que sigue es

intentar superar esas viejas y tediosas categorías. Obsérvese que nadie tiene la sensación de que tenga que elegir entre monismo y dualismo cuando lo que está en juego es el problema 'digestión-estómago'. ¿Por qué ha de suceder algo diferente con el problema 'mente-cuerpo'?

Pero, vocabulario aparte, hay aún un problema o familia de problemas. Desde Descartes, el problema mente-cuerpo ha tomado la forma siguiente: ¿Cómo podemos dar cuenta de las relaciones entre dos géneros de cosas, en apariencia totalmente diferentes? Por un lado hay cosas materiales, tales como nuestros pensamientos y sensaciones: pensamos de ellos que son subjetivos, conscientes e inmateriales. Por otro lado, hay cosas físicas; pensamos de ellas que tienen una masa, que se extienden en el espacio y que interactúan causalmente con otras cosas físicas. La mayor parte de las soluciones intentadas al problema mente-cuerpo concluyen negando la existencia de, o degradando de algún modo el estatus de, uno u otro de esos tipos de cosa. Dado el éxito de las ciencias físicas no es sorprendente que en nuestro estadio de desarrollo intelectual la tentación sea degradar el estatus de las entidades mentales. Así, la mayor parte de las concepciones materialistas de la mente, actualmente en boga —tales como el conductismo, el funcionalismo y el fisicalismo— terminan negando implícita o explícitamente que haya cosas tales como las mentes del modo en que las pensamos ordinariamente. Esto es, niegan que, en realidad, tengamos *intrínsecamente* estados subjetivos, conscientes mentales, y que sean tan reales y tan irreductibles como cualquier cosa del universo.

Ahora bien, ¿por qué hacen esto? ¿Por qué sucede que tantos teóricos terminen por negar el carácter intrínsecamente mental de los fenómenos mentales? Si podemos responder a esta cuestión creo que entenderemos por qué el problema mente-cuerpo ha parecido tan intratable durante largo tiempo.

Hay cuatro rasgos de los fenómenos mentales que han hecho que parezcan imposibles de encajar dentro de nuestra concepción 'científica' del mundo como compuesto de cosas materiales. Y son esos cuatro rasgos los que

han hecho realmente difícil el problema mente-cuerpo. Son tan embarazosas que han llevado a muchos pensadores en filosofía, psicología e inteligencia artificial, a decir cosas extrañas e implausibles sobre la mente.

El más importante de esos rasgos es la conciencia. Yo, en el momento de escribir esto, y usted, en el momento de leerlo, somos ambos conscientes. Es justamente un hecho puro y simple sobre el mundo el que éste contiene tales estados y eventos mentales conscientes, pero es difícil ver cómo sistemas meramente físicos pueden tener conciencia. ¿Cómo puede ocurrir tal cosa? ¿Cómo, por ejemplo, puede esa masa informe gris y blanca que está dentro de mi cráneo ser consciente?

Concibo que la existencia de la conciencia pueda parecernos asombrosa. Es bastante fácil imaginarnos un universo sin ella, pero si se hace se verá que es un universo que verdaderamente carece de significado. La conciencia es el hecho central de la existencia específicamente humana, puesto que sin ella todos los demás aspectos específicamente humanos de nuestra existencia —lenguaje, amor, humor y así sucesivamente— serían imposibles. Creo, dicho sea de paso, que tiene algo de escándalo el que las discusiones contemporáneas en filosofía y en psicología tengan tan poco interés en hablarnos sobre la conciencia.

El segundo rasgo intratable de la mente es lo que los filósofos y los psicólogos llaman 'intencionalidad' el rasgo mediante el cual nuestros estados mentales se dirigen a, o son sobre, o se refieren a, o son de objetos y estados de cosas del mundo distintos de ellos mismos. 'Intencionalidad', dicha sea de paso, no se refiere sólo a intensiones, sino también a creencias, deseos, esperanzas, temores, amor, odio, lascivia, aversión, vergüenza, orgullo, irritación, diversión, y todos aquellos estados mentales (conscientes o inconscientes) que se refieren a, o son sobre, el mundo distinto de la mente. Ahora bien, la cuestión sobre la intencionalidad es muy parecida a la cuestión sobre la conciencia. ¿Cómo puede esta materia que hay dentro de mi cabeza ser *sobre* algo? ¿Cómo se puede *referir* a algo? Después de todo, esta materia que hay dentro del cráneo consta de 'áto-

mos en el vacío', lo mismo que el resto de la realidad material consta de átomos en el vacío. Ahora bien, para decirlo crudamente, ¿cómo pueden los átomos en el vacío representar algo?

El tercer rasgo de la mente que parece difícil de acomodar dentro de una concepción científica de la realidad es la subjetividad de los estados mentales. Esta subjetividad está marcada por hechos tales como que yo puedo sentir mis dolores y tú no puedes. Yo veo el mundo desde mi punto de vista, tú lo ves desde tu punto de vista. Yo soy consciente de mí mismo y de mis estados mentales internos, como algo completamente distinto de los yoes y los estados mentales de otras personas. Desde el siglo XVII hemos dado en pensar que la realidad es algo que tiene que ser igualmente accesible a todos los observadores competentes; esto es, pensamos que tiene que ser objetiva. Ahora bien, ¿cómo hemos de acomodar la realidad de los fenómenos mentales *subjetivos* con la concepción científica de la realidad como algo totalmente *objetivo?*

Finalmente, hay un cuarto problema, el problema de la causación mental. Todos nosotros suponemos, como parte del sentido común, que nuestros pensamientos y sensaciones tienen realmente importancia para el modo en que nos comportamos, que tienen de hecho algún efecto *causal* sobre el mundo físico. Yo decido, por ejemplo, levantar mi brazo y —he aquí— mi brazo se levanta. Pero si nuestros pensamientos y sensaciones son verdaderamente mentales, ¿cómo pueden afectar a algo físico? ¿Cómo puede algo mental tener una influencia física? ¿Se supone que pensamos que nuestros pensamientos y sensaciones pueden producir de algún modo efectos químicos sobre nuestros cerebros y el resto de nuestro sistema nervioso? ¿Cómo podría ocurrir tal cosa? ¿Se supone que pensamos que los pensamientos pueden enroscarse alrededor de los axones o sacudir las dentritas, o colarse dentro de la membrana celular y atacar el núcleo de la célula?

Pero a menos que tenga lugar alguna conexión de este tipo entre la mente y el cerebro, ¿no quedamos solamente con el punto de vista de que la mente no impor-

ta, que es tan poco importante causalmente como la espuma de la ola lo que es para su movimiento? Supongo que si la espuma fuese consciente, podría pensar de sí mismo: '¡Qué tarea tan ardua es tirar de estas olas hacia la playa y luego hacerlas retroceder de nuevo, y así todo el día!' Pero sabemos que la espuma no tiene realmente ninguna importancia significativa. ¿Por qué suponemos que nuestra vida mental es algo más importante que una porción de espuma en la ola de la realidad física?

Estos cuatro rasgos, conciencia, intencionalidad, subjetividad y causación mental son los que hacen que el problema mente-cuerpo parezca tan difícil. Con todo. quiero decir que todos ellos son rasgos reales de nuestras vidas mentales. No todo estado mental los posee todos. Pero cualquier explicación satisfactoria de la mente y de las relaciones mente-cuerpo tiene que tomar en cuenta la totalidad de los cuatro rasgos. Si una teoría acaba negando cualquiera de ellos, con ello se sabe que se tiene que haber cometido algún error en algún lugar.

La primera tesis que quiero abrazar con vistas a 'resolver el problema mente-cuerpo' es ésta:

> *Los fenómenos mentales, todos los fenómenos mentales, ya sean conscientes o inconscientes, visuales o auditivos, dolores, cosquilleos, picazones, pensamientos, toda nuestra vida mental, están efectivamente causados por procesos que acaecen el cerebro.*

Para tener una impresión de cómo funciona esto, intentemos describir los procesos causales con algún detalle para, al menos, un género de estado mental. Consideremos, por ejemplo, los dolores. Desde luego, cualquier cosa que digamos ahora puede parecer maravillosamente pintoresca dentro de una generación, a medida que se fundamente nuestro conocimiento de cómo funciona el cerebro. Con todo, la *forma* de la explicación puede seguir siendo válida aunque los *detalles* se alteren. De acuerdo con los puntos de vista actuales, las señales de dolor se transmiten desde las terminaciones sensoriales nerviosas a la médula espinal por medio de al

menos dos tipos de fibras —hay las fibras A Delta, que están especializadas en sensaciones punzantes, y las fibras C, que están especializadas en sensaciones de quemazón, y en aquellas que implican un dolor continuado. En la médula espinal pasan a través de una región denominada el tracto de Lissauer y terminan en las neuronas de la médula. A medida en que las señales suben por la medula espinal entran en el cerebro por dos caminos separados: el camino del dolor punzante y el camino del dolor producido por quemazón. Ambos caminos pasan a través del tálamo, pero el dolor punzante está más localizado más adelante, en el córtex somatosensorial, mientras que el camino del dolor producido por quemazón transmite señales no sólo hacia arriba, hacia el interior del córtex, sino también lateralmente, hacia el hipotálamo y otras regiones de la base del cerebro. A causa de esas diferencias es mucho más fácil para nosotros localizar una sensación punzante —podemos decir bastante exactamente cuándo alguien está clavando un alfiler en nuestra piel, por ejemplo—, mientras que los dolores producidos por quemazón o los continuos pueden ser más penosos, puesto que activan más el sistema nervioso. La sensación efectiva de dolor parece, estar causada tanto por la estimulación de las regiones basales del cerebro, especialmente el tálamo, y la estimulación del córtex somatosensorial.

Ahora bien, para los propósitos de esta discusión, el punto que necesitamos remachar es éste: nuestras sensaciones de dolor están causadas por una serie de eventos que comienzan en las terminaciones nerviosas libres y terminan en el tálamo y en otras regiones del cerebro. De hecho, por lo que respecta a las sensaciones efectivas, los eventos que acaecen dentro del sistema nervioso central son completamente suficientes para causar dolores —nosotros sabemos esto tanto por los dolores de los miembros-fantasma que sienten los amputados, como por los dolores causados estimulando artificialmente porciones relevantes del cerebro. Quiero sugerir que lo que es verdadero del dolor es verdadero generalmente de los fenómenos mentales. Para decirlo crudamente, y considerando para nuestra presente discusión todo el sistema ner-

vioso central como parte del cerebro: todos nuestros pensamientos y sensaciones están causados por procesos que ocurren dentro del cerebro. Por lo que respecta a la causa de los estados mentales, el paso crucial es el que tiene lugar dentro de la cabeza, no el estímulo externo o periférico. Y la argumentación a favor de esto es simple. Si ocurrieran los eventos de fuera del sistema nervioso central, pero no sucediese nada en el cerebro, entonces no habría eventos mentales. Pero si ocurriesen las cosas correctas en el cerebro, incluso si no hubiese estímulos externos, los eventos mentales ocurrirían. (Y esto, dicho sea de paso, es el principio según el cual trabaja la anestesia quirúrgica: se impide que el estímulo externo tenga los efectos relevantes sobre el sistema nervioso central.)

Pero si los dolores y otros fenómenos mentales están causados por procesos que tienen lugar en el cerebro, se querrá saber ¿qué son los dolores? ¿Qué son realmente? Bien, en el caso de los dolores, la respuesta obvia es que son tipos de sensaciones desagradables. Pero esa respuesta nos deja insatisfechos porque no nos dice cómo encajan los dolores en nuestra concepción global del mundo.

Una vez más pienso que la respuesta a esta cuestión es obvia, pero nos costará un poco de trabajo el escribirla con todas sus letras. A nuestra primera afirmación, que los dolores y otros fenómenos mentales están causados por procesos cerebrales, necesitamos añadir una segunda afirmación.

> *Los dolores y otros fenómenos mentales son sólo rasgos del cerebro (y quizá del resto del sistema nervioso central).*

Una de las aspiraciones primarias de este capítulo es mostrar cómo *ambas* proposiciones pueden ser verdaderas a la vez. ¿Cómo puede ser el caso de que los cerebros causen las mentes, y, con todo, que las mentes sean sólo rasgos de los cerebros? Creo que el no lograr ver cómo estas dos proposiciones pueden ser a la vez verdaderas es lo que ha bloqueado durante tanto tiempo una solución al problema mente-cuerpo. Hay diferentes niveles

de confusión que tal par de ideas puede generar. Si los fenómenos mentales y físicos tienen relaciones de causa y efecto, ¿cómo puede ser uno un rasgo de otro? ¿No implicaría esto que la mente se ha causado a sí misma, la espantosa doctrina de la *causa sui?* Pero en el fondo de esta perplejidad está una mala comprensión de la causación. Es tentador pensar que siempre que A causa B tiene que haber dos eventos discretos, uno identificado como la causa, el otro identificado como el efecto; que toda causación funciona del mismo modo que las bolas de billar golpeándose entre sí. Este crudo modelo de las relaciones causales entre el cerebro y la mente nos inclina a aceptar algún género de dualismo; estamos inclinados a pensar que los eventos de un reino material, el 'físico', causan eventos en otro reino insubstancial, el 'mental'. Pero esto me parece un error. Y una manera de eliminar el error es lograr un concepto de causación más sofisticado. Para hacer esto me apartaré por un momento de las relaciones entre mente y cerebro para observar algunos otros tipos de relaciones causales que se dan en la naturaleza.

Una distinción común en física es aquella que se da entre micro y macropropiedades de sistemas a pequeña y a gran escala. Considérese, por ejemplo, la mesa a la que estoy sentado ahora, o el vaso de agua que está delante de mí. Cada objeto está compuesto de micropartículas. Las micropartículas tienen rasgos al nivel de las moléculas y átomos, así como la solidez de la mesa, la liquidez del agua y la transparencia del vaso, que son rasgos superficiales o globales de los sistemas físicos. Muchas de esas propiedades superficiales o globales pueden explicarse causalmente por la conducta de elementos del micronivel. Por ejemplo, la solidez de la mesa que está delante de mí se explica por la estructura de enrejado ocupada por las moléculas de las que está compuesta. Similarmente, la liquidez del agua se explica por la naturaleza de las interacciones entre las moléculas de H_2O. Esos macrorrasgos se explica causalmente por la conducta de elementos de micronivel.

Quiero sugerir que esto proporciona un modelo per-

fectamente ordinario para explicar las problemáticas relaciones entre la mente y el cerebro. En el caso de la solidez, liquidez y transparencia, no tenemos dificultad alguna en suponer que los rasgos superficiales son *causados por* la conducta de elementos del micronivel, y al mismo tiempo aceptamos que los fenómenos superficiales *son solamente* rasgos de los mismos sistemas en cuestión. Pienso que el modo más clásico de enunciar este punto, es decir, que el rasgo superficial es *causado por* la conducta de los microelementos, y al mismo tiempo está *realizado en* el sistema que está compuesto de los microelementos. Hay una relación de causa y efecto, pero al mismo tiempo los rasgos superficiales son sólo rasgos de nivel superior del mismo sistema cuyo comportamiento en el micronivel causa esos rasgos.

Alguien podría decir, como objeción a esto, que la liquidez, la solidez y así sucesivamente son idénticas a los rasgos de la microestructura. Así, por ejemplo, podríamos definir solidez como la estructura enrejada de la disposición molecular, lo mismo que el calor se identifica a menudo con la energía cinética media de los movimientos de las moléculas. Este punto me parece correcto, pero no me parece realmente una objeción al análisis que estoy proponiendo. Es una característica del progreso de la ciencia el que una expresión que se define originalmente en términos de rasgos superficiales, rasgos accesibles a los sentidos, sea subsecuentemente definida en términos de la microestructura que causa esos rasgos superficiales. Así, para tomar el ejemplo de la solidez, la mesa que está enfrente de mí es sólida en el sentido ordinario de que es rígida, es resistente a la presión, sostiene libros, no es fácilmente penetrable por la mayor parte de otros objetos, tales como otras mesas, y así sucesivamente. Tal es la noción de sentido común de solidez. Y en vena científica se puede definir solidez como cualquier microestructura que causa esos rasgos observables. Así puede decirse o que la solidez es la estructura enrejada del sistema de moléculas y que la salidez así definida causa, por ejemplo, la resistencia al tacto y a la presión. O puede decirse que la solidez consiste en rasgos de nivel superior, tales como la rigidez y

la resistencia al tacto y a la presión y que es causada por la conducta de los elementos del micronivel.

Si aplicamos estas lecciones al estudio de la mente, me parece que no hay dificultad en dar cuenta de las relaciones de la mente con el cerebro en términos del funcionamiento del cerebro para causar estados mentales. Lo mismo que la liquidez del agua es causada por la conducta de elementos del micronivel y, con todo, es al mismo tiempo un rasgo realizado en el sistema de microelementos, así exactamente en ese sentido de 'causado por' y 'realizado en', los fenómenos mentales son causados por procesos que tienen lugar en el cerebro en el nivel neuronal o modular, y al mismo tiempo se realizan en el sistema mismo que consta de neuronas. Y lo mismo que necesitamos la distinción micro-macro para cualquier sistema físico, así también, por las mismas razones, necesitamos la distinción micro-macro para el cerebro. Y aunque podamos decir de un sistema de partículas que está a 10° C. o que es sólido o que es líquido, no podemos decir de ninguna partícula dada que esa partícula es sólida, que esa partícula es líquida o que esa partícula está a 10° C. No puedo, por ejemplo, meter la mano dentro de este vaso de agua, sacar una molécula y decir: 'Esta molécula está mojada.'

De la misma manera, en la medida en que sabemos algo sobre ello, aunque podemos decir de un cerebro particular: 'Este cerebro es consciente' o 'este cerebro está teniendo una experiencia de sed o de dolor' no podemos decir de ninguna neurona particular del cerebro: 'Esta neurona tiene dolor, esta neurona está teniendo una experiencia de sed.' Para repetir este punto, aunque hay enormes misterios empíricos sobre cómo funciona con detalle el cerebro, no hay obstáculos lógicos, o filosóficos, o metafísicos, para dar cuenta de la relación entre la mente y el cerebro en términos que son completamente familiares para nosotros a partir del resto de la naturaleza. Nada hay más común en la naturaleza que el que rasgos superficiales de un fenómeno sean a la vez causados por y realizados en una microestructura, y ésas son exactamente las relaciones que se exhiben en la relación de la mente con el cerebro.

Volvamos a los cuatro problemas a los que he dicho que se enfrentaba cualquier intento de solucionar el problema mente-cerebro. Primero, ¿cómo es posible la conciencia?

La mejor manera de mostrar cómo algo es posible es mostrar que existe efectivamente. Ya hemos hecho un bosquejo de cómo los dolores son causados por procesos neurofisiológicos que se desarrollan en el tálamo y en el córtex sensorial. ¿Por qué entonces tanta gente se siente insatisfecha con este tipo de respuesta? Pienso que siguiendo la pista de una analogía con un problema anterior de la historia de la ciencia podemos disipar esa sensación de perplejidad. Durante mucho tiempo muchos biólogos y filósofos pensaron que era imposible, en principio, dar cuenta de la existencia de la *vida* de acuerdo con fundamentos puramente biológicos. Pensaban que además de los procesos biológicos tenía que ser necesario algún otro elemento, tenía que postularse algún *élan vital* para dar vida a lo que era de otra manera materia muerta e inerte. Es difícil hoy en día darse cuenta de cuán intensa fue la disputa entre vitalismo y mecanicismo, no hace ni siquiera una generación; pero hoy en día esas cuestiones ya no se toman en serio. ¿Por qué no? Pienso que no es tanto porque el mecanicismo ganase y el vitalismo perdiese, sino porque hemos llegado a entender mejor el carácter biológico de los procesos que son característicos de los organismos vivientes. Una vez que entendemos cómo los rasgos que son característicos de los seres vivientes tienen una explicación biológica, ya no nos parece misterioso el que la materia esté viva. Pienso que consideraciones exactamente similares se han de aplicar a nuestra discusión de la conciencia. En principio, ya no debe parecer nada misterioso que este hermoso trozo de materia, esta substancia gris y blanca de textura similar a la de la harina de avena, sea consciente, que este otro hermoso trozo de materia, esta colección de moléculas nucleoproteicas pegadas alrededor de un armazón de calcio, esté vivo. Dicho brevemente, la manera de disipar el misterio es entender el proceso. Ciertamente todavía no entendemos completamente el proceso, pero entendemos su *carácter* general, entendemos

que hay ciertas actividades electroquímicas específicas que se desarrollan entre las neuronas o los módulos de las neuronas y quizá otros rasgos del cerebro, y esos procesos causan la conciencia.

Nuestro segundo problema era: ¿cómo pueden los átomos en el vacío tener intencionalidad? ¿Cómo pueden ser sobre algo?

Como sucedía con nuestra primera cuestión, la mejor manera de mostrar cómo algo es posible, es mostrar cómo efectivamente existe. Así, consideremos la sed. En la medida en que conocemos algo sobre esto, al menos ciertos géneros de sed son causados en el hipotálamo por secuencias de activaciones nerviosas. Esas activaciones son causadas a su vez por la acción de la angiotensina en el hipotálamo, y la angiotensina, a su vez, es sintetizada por la renina, que es segregada por los riñones. La sed, al menos de esos géneros, es causada por una serie de eventos en el sistema nervioso central, principalmente en el hipotálamo, y tiene su realización en el hipotálamo. Estar sediento es tener, entre otras cosas, el deseo de beber. La sed es, por consiguiente, un estado intencional: tiene contenido; su contenido determina bajo qué condiciones es satisfecha, y tiene todas las demás características que son comunes a los estados intencionales.

Como sucede con los 'misterios' de la vida y de la conciencia, la manera de dominar el misterio de la intencionalidad es describir con tanto detalle como podamos cómo los fenómenos son causados por procesos biológicos mientras que al mismo tiempo tienen su realización en los sistemas biológicos. Las experiencias visuales y auditivas, las sensaciones táctiles, el hambre, la sed y el deseo sexual son causados todos ellos por procesos cerebrales y tienen su realización en la estructura del cerebro, y todos ellos son fenómenos intencionales.

No estoy diciendo que debamos perder nuestro sentido de los misterios de la naturaleza. Al contrario, los ejemplos que he citado son todos ellos asombrosos en un sentido. Pero estoy diciendo que no son ni más ni menos misteriosos que otras asombrosas características del mundo, tales como la existencia de la atracción gravitatoria, el proceso de la fotosíntesis o el tamaño de la Vía Láctea.

Nuestro tercer problema era ¿cómo acomodar la subjetividad de los estados mentales dentro de una concepción objetiva del mundo real?

Me parece un error suponer que la definición de realidad deba excluir la subjetividad. Si 'ciencia' es el nombre de la colección de verdades objetivas y sistemáticas que podemos enunciar sobre el mundo, entonces la existencia de la subjetividad es un hecho científico objetivo igual que cualquier otro. Si una explicación científica del mundo intenta describir cómo son las cosas, entonces uno de los rasgos de la explicación será la subjetividad de los estados mentales, puesto que un hecho puro y simple de la evolución biológica es que ha producido ciertos tipos de sistemas biológicos, a saber: los cerebros humanos y los de ciertos animales, que tienen rasgos subjetivos. Mi actual estado de conciencia es un rasgo de mi cerebro, pero sus aspectos conscientes son accesibles para mí de una manera en que no son accesibles para usted. Y su estado actual de conciencia es un rasgo de su cerebro y sus aspectos conscientes son accesibles para usted de una manera en que no son accesibles para mí. Así pues, la existencia de la subjetividad es un hecho objetivo de la biología. Es un error persistente el intentar definir 'ciencia' en términos de ciertos rasgos de teorías científicas existentes. Pero una vez que se percibe lo perjudicial que es este provincialismo, entonces cualquier dominio de hechos es un tema de investigación sistemática. Así, por ejemplo, si Dios existiese, este hecho sería un hecho igual que cualquier otro. No sé si Dios existe, pero no tengo duda alguna de que todos los estados mentales subjetivos existen, puesto que yo estoy ahora en uno de ellos, lo mismo que lo está usted. Si el hecho de la subjetividad va en contra de cierta definición de 'ciencia', entonces lo que hemos de abandonar es la definición y no el hecho.

En cuarto lugar, el problema de la causación mental es, para nuestros presentes propósitos, explicar cómo los eventos mentales pueden causar eventos físicos. ¿Cómo, por ejemplo, puede algo tan 'carente de peso' y tan 'etéreo' como el pensamiento dar lugar a una acción?

La respuesta es que los pensamientos ni carecen de peso ni son etéreos. Cuando se tiene un pensamiento, se

está desarrollando actividad cerebral. La actividad cerebral causa movimientos corporales por medio de los procesos fisiológicos. Ahora bien, puesto que los estados mentales son rasgos del cerebro, tienen dos niveles de descripción: un nivel superior en términos mentales y un nivel inferior en términos fisiológicos. Los mismos poderes causales del sistema pueden ser descritos a cualquiera de los dos niveles.

Una vez más podemos considerar una analogía tomada de la física para ilustrar esas relaciones. Considérese el martillear un clavo con un martillo. Tanto el martillo como el clavo tienen un cierto género de solidez. Los martillos hechos de algodón en rama o de mantequilla son completamente inútiles, y los martillos hechos de agua o de vapor no son martillos en absoluto. La solidez es una propiedad causal real del martillo. Pero la solidez misma está causada por la conducta de las partículas del micronivel y está realizada en el sistema que consta de microelementos. La existencia en el cerebro de dos niveles de descripción causalmente reales, uno un macronivel de procesos mentales y el otro un micronivel de procesos neuronales, es exactamente análogo a la existencia de dos niveles causalmente reales de descripción del martillo. La conciencia, por ejemplo, es una propiedad real del cerebro que causa que las cosas sucedan. Mi intento consciente de realizar una acción, tal como levantar mi brazo, causa el movimiento del brazo. En el nivel superior de descripción, la intención de levantar mi brazo causa el movimiento del brazo. Pero en el nivel inferior de descripción, una serie de activaciones neuronales comienza una cadena de eventos que da como resultado la contracción de los músculos. Como sucede con el caso de martillear un clavo, la misma secuencia de eventos tiene dos niveles de descripción. Ambos son causalmente reales, y los rasgos causales del nivel superior están a la vez causados por y realizados en la estructura de los elementos de nivel inferior.

Resumamos: según mi punto de vista, la mente y el cuerpo interactúan, pero no son dos cosas diferentes, puesto que los fenómenos mentales son solamente rasgos del cerebro. Una manera de caracterizar esta posi-

ción es verla como una aserción de, a la vez, fisicalismo y mentalismo. Supóngase que definimos fisicalismo 'ingenuo' como el punto de vista de que todo lo que existe en el mundo son partículas físicas con sus propiedades y relaciones. La potencia del modelo físico de la realidad es tan grande que es difícil ver cómo podemos desafiar seriamente el fisicalismo ingenuo. Y definamos 'mentalismo ingenuo' como el punto de vista de que los fenómenos mentales existen realmente. Hay realmente estados mentales; algunos de ellos son conscientes; muchos tienen intencionalidad; todos ellos tienen subjetividad; y muchos de ellos funcionan casualmente determinando eventos físicos del mundo. La tesis de este primer capítulo puede ahora enunciarse muy simplemente. El mentalismo ingenuo y el fisicalismo ingenuo son perfectamente coherentes entre sí. Es más, en la medida en que sabemos algo sobre cómo funciona el mundo, no son solamente coherentes; ambos son verdaderos.

2

¿PUEDEN LOS COMPUTADORES PENSAR?

En el capítulo anterior he proporcionado, al menos, las líneas generales de una solución al llamado 'problema mente-cuerpo'. Aunque no sabemos con detalle cómo funciona el cerebro, sabemos lo suficiente para tener una idea de las relaciones generales entre los procesos cerebrales y los procesos mentales. Los procesos mentales están causados por la conducta de elementos del cerebro. Al mismo tiempo, se realizan en la estructura que está compuesta por esos elementos. Pienso que esta respuesta es coherente con los enfoques biológicos estándar de los fenómenos biológicos. De hecho, es un género de respuesta de sentido común a la cuestión, dado lo que sabemos acerca de cómo el mundo funciona. Sin embargo, es, con mucho, el punto de vista de una minoría. El punto de vista prevalece en filosofía, psicología e inteligencia artifical es aquél que subraya las analogías entre el funcionamiento del cerebro y el funcionamiento de los computadores digitales. De acuerdo con la versión más extrema de este punto de vista, el cerebro es solamente un computador digital y la mente es solamente un programa de computador. Podría resumirse este punto de vista; yo lo llamo 'inteligencia artificial fuerte', o 'IA fuerte', diciendo que la mente es al cerebro lo que el programa es al *hardware* del computador.

Este punto de vista tiene la consecuencia de que no hay nada esencialmente biológico por lo que respecta a la mente humana. Sucede solamente que el cerebro es uno de un número indefinidamente extenso de diferentes géneros de *hardware* de computador que podrían servir de sostén a los programas que constituyen la inteligencia humana. Según este punto de vista, cualquier sistema físico que tuviese el programa correcto con los *inputs* y los *outputs* correctos tendría una mente en exactamente el mismo sentido que tú y yo tenemos mentes. Así, por ejemplo, si se hiciese un computador con viejas latas de cerveza y se le suministrase energía por medio de molinillos de viento, si tuviera el programa correcto, tendría que tener una mente. Y el punto no es que, dado todo lo que sabemos, podría tener pensamientos y sensaciones, sino más bien que tiene que tener pensamientos y sensaciones, puesto que todo aquello en lo que consiste tener pensamientos y sensaciones es esto: llevar a cabo el programa correcto.

La mayor parte de la gente que mantiene este punto de vista piensa que todavía no hemos diseñado programas que sean mentes. Pero hay bastante acuerdo general entre ellos de que esto es solamente un asunto de tiempo hasta que los científicos computacionales y las personas que trabajan en inteligencia artifical diseñen el *hardware* y los programas apropiados que sean los equivalentes de los cerebros y mentes humanas. Estos serán cerebros y mentes artificiales que son en todos los sentidos los equivalentes de los cerebros y las mentes humanos.

Mucha gente que está fuera del campo de la inteligencia artificial queda completamente pasmada al descubrir que alguien pueda creer un punto de vista como éste. Así pues, antes de criticarlo, permítaseme dar un puñado de ejemplos de las cosas que la gente que trabaja en este campo ha dicho efectivamente. Herbert Simon, de la Universidad de Carnegie-Mellon, dice que ya tenemos máquinas que literalmente pueden pensar. Ya no es cuestión de esperar por ninguna máquina futura, puesto que existen ya computadores digitales que

tienen pensamientos exactamente en el mismo sentido que usted y yo los tenemos. Bien, ¡qué casualidad! Los filósofos han estado preocupados durante siglos por la cuestión de si una máquina podría o no pensar, y ahora descubrimos que en Carnegie-Mellon ya tienen esas máquinas. El colega de Simon, Alan Newell, afirma que hemos descubierto ahora (y obsérvese que Newell dice 'descubierto' y no 'hemos avanzado la hipótesis' o 'hemos descubierto la posibilidad', sino hemos *descubierto*) que la inteligencia es solamente un asunto de manipulación de símbolos físicos; no tiene ninguna conexión esencial con ningún género específico de *wetware* o *hardware* biológico o físico. Más bien, cualquier sistema que sea capaz de manipular símbolos físicos de una manera correcta es capaz de inteligencia en el mismo sentido literal que la inteligencia humana de los seres humanos. Tanto Simon como Newell subrayan que no hay nada metafórico en esas afirmaciones; las proponen de una manera completamente literal. Se cita a Freeman Dyson como el que dijo que los computadores tienen una ventaja sobre el resto de nosotros por lo que respecta a la evolución. Puesto que la conciencia es un asunto de procesos formales solamente, en los computadores esos procesos formales pueden tener lugar en substancias que son mucho más capaces de sobrevivir en un universo que está enfriando, que en seres como nosotros, hechos de nuestros húmedos y sucios materiales. Marvin Minsky, del MIT, dice que la próxima generación de computadores será tan inteligente que deberíamos 'estar contentos si estuvieran dispuestos a mantenernos en torno a la casa como animalitos domésticos'. Mi siempre favorito en la literatura de afirmaciones exageradas a favor de los computadores es John McCarthy, el inventor del término 'inteligencia artificial'. MacCarthy dice que incluso 'puede decirse que máquinas tan simples como los termostatos tienen creencias'. Y de hecho, de acuerdo con él, de toda máquina capaz de resolver problemas puede decirse que tiene creencias. Admiro el coraje de MacCarthy. Una vez le pregunté. '¿Qué creencias tiene su termostato?' Y él me dijo: 'Mi termostato

tiene tres creencias: hace demasiado calor aquí, hace demasiado frío aquí, y aquí hace la temperatura correcta'. Como filósofo me gustan esas tres afirmaciones por una simple razón. A diferencia de muchas tesis filosóficas, son razonablemente claras, y admiten una refutación simple y decisiva. Es esta refutación la que voy a emprender en el presente capítulo.

La naturaleza de la refutación no tiene nada que ver con ninguna etapa particular de la tecnología de los computadores. Es importante subrayar este punto, puesto que la tentación es siempre pensar que la solución a nuestros problemas tiene que esperar a alguna, hasta ahora no creada, maravilla tecnológica. Pero de hecho, la naturaleza de la refutación es completamente independiente de cualquier estado en que se encuentre la tecnología. No tiene nada que ver con la definición misma de computador digital, con lo que un computador digital es.

Es esencial para nuestra concepción de computador digital que sus operaciones puedan especificarse de manera completamente formal; esto es, nosotros especificamos los pasos de la operación del computador en términos de símbolos abstractos —secuencias de ceros y unos impresos en una cinta, por ejemplo. Una 'regla típica de computador determinará que cuando una máquina está en un cierto estado y tiene un cierto símbolo en su cinta, entonces realizará ciertas operaciones tales como borrar el símbolo o escribir otro símbolo y a continuación entrar en otro estado tal como mover la cinta un cuadrado a la izquierda. Pero los símbolos no tienen ningún significado, no tienen ningún contenido semántico, no se refieren a nada. Tienen que especificarse en términos puramente de su estructura formal o semántica. Los ceros y los unos, por ejemplo, son solamente numerales, no están ni siquiera por números. Es más, es esta característica de los computadores digitales la que los hace tan potentes. Uno y el mismo tipo de *hardware*, si se diseña apropiadamente, puede usarse para pasar un rango indefinido de programas diferentes. Y uno, y el

mismo programa puede ser pasado en un rango indefinido de diferentes tipos de *hardware.*

Pero este rasgo de los programas, el que estén definidos de manera puramente formal o sintáctica, es fatal para el punto de vista de que los procesos mentales y los procesos de programas son idénticos. Y la razón puede enunciarse de manera completamente simple. Tener una mente es algo más que tener procesos formales o sintácticos. Nuestros estados mentales internos tienen, por definición, ciertos tipos de contenido. Si estoy pensando en Kansas City, o deseando tener una cerveza fría para beber, o preguntándome si habrá una caída en los tipos de interés, en cada caso mi estado mental tiene un cierto contenido mental además de cualesquiera otros rasgos formales que pueda tener. Esto es, incluso si mis pensamientos se me presentan en cadenas de símbolos tiene que haber más que las cadenas abstractas, puesto que las cadenas por sí mismas no pueden tener significado alguno. Si mis pensamientos han de ser *sobre* algo, entonces las cadenas tienen que tener un *significado* que hace que sean los pensamientos sobre esas cosas. En una palabra, la mente tiene más que una sintaxis, tiene una semántica. La razón por la que un programa de computador no pueda jamás ser una mente es simplemente que un programa de computador es solamente sintáctico, y las mentes son más que sintácticas. Las mentes son semánticas, en el sentido de que tienen algo más que una estructura formal: tienen un contenido.

Para ilustrar este punto he diseñado un cierto experimento de pensamiento. Imaginemos que un grupo de programadores de computador ha escrito un programa que capacita a un computador para simular que entiende chino. Así, por ejemplo, si al computador se le hace una pregunta en chino, confrontará la pregunta con su memoria o su base de datos, y producirá respuestas adecuadas a las preguntas en chino. Supongamos, por mor del argumento, que las respuestas del computador son tan buenas como las de un hablante nativo del chino. Ahora bien, ¿entiende el computador, según esto, chino? ¿Entiende literalmente chino, de la manera en que los hablantes del chino entienden chino? Bien, imaginemos

que se le encierra a usted en una habitación y que en esta habitación hay diversas cestas llenas de símbolos chinos. Imaginemos que usted (como yo) no entiende chino, pero que se le da un libro de reglas en castellano para manipular esos símbolos chinos. Las reglas especifican las manipulaciones de los símbolos de manera puramente formal, en términos de su sintaxis, no de su semántica. Así la regla podría decir: 'toma un signo changyuan-changyuan de la cesta número uno y ponlo al lado de un signo chongyuon-chongyuon de la cesta número dos'. Supongamos ahora que son introducidos en la habitación algunos otros símbolos chinos, y que se le dan reglas adicionalen para devolver símbolos chinos fuera de la habitación. Supóngase que usted no sabe que los símbolos introducidos en la habitación son denominados 'preguntas' de la gente que está fuera de la habitación, y que los símbolos que usted devuelve fuera de la habitación son denominados 'respuestas a las preguntas'. Supóngase, además, que los programadores son tan buenos al diseñar los programas y que usted es tan bueno manipulando los símbolos que enseguida sus respuestas son indistinguibles de las de un hablante nativo del chino. He aquí que usted está encerrado en su habitación barajando sus símbolos chinos y devolviendo símbolos chinos en respuesta los símbolos chinos que entran. Sobre la base de la situación tal como la he descrito, no hay manera de que usted pueda aprender nada de chino manipulando esos símbolos formales.

Ahora bien, lo esencial de la historieta es simplemente esto: en virtud del cumplimiento de un programa de computador formal desde el punto de vista de un observador externo, usted se comporta exactamente como si entendiese chino, pero a pesar de todo usted no entiende ni palabra de chino. Pero si pasar por el programa de computador apropiado para entender chino no es suficiente para proporcionarle a *usted* comprensión del chino, entonces no es suficiente para proporcionar a *cualquier otro computador digital* comprensión del chino. Y nuevamente, la razón de esto puede enunciarse muy simplemente. Si usted no entiende chino, entonces ningún otro computador podría entender chino puesto

que ningún computador digital, en virtud solamente de pasar un programa, tiene nada que usted no tenga. Todo lo que el computador tiene, como usted tiene también, es un programa formal para manipular símbolos chinos no interpretados. Para repetirlo: un computador tiene una sintaxis, pero no una semántica. Todo el objeto de la parábola de la habitación china es recordarnos un hecho que conocíamos desde el principio. Comprender un lenguaje, o ciertamente, tener estados mentales, incluye algo más que tener un puñado de símbolos formales. Incluye tener una interpretación o un significado agregado a esos símbolos. Y un computador digital, tal como se ha definido, no puede tener más que símbolos formales puesto que la operación del computador, como dije anteriormente, se define en términos de su capacidad para llevar a cabo programas. Y esos programas son especificables de manera puramente formal —esto es, no tienen contenido semántico.

Podemos ver la fuerza de este argumento si contrastamos aquello a lo que se parece el ser preguntado y responder a preguntas en algún lenguaje en el que no tenemos conocimiento alguno de ninguno de los significados de las palabras. Imaginemos que en la habitación china se le dan también a usted preguntas en castellano sobre cosas tales como su edad o episodios de su vida, y que usted responde a esas preguntas. ¿Cuál es la diferencia entre el caso del chino y el caso del castellano? Bien, si igual que yo usted no entiende nada de chino y entiende castellano, entonces la diferencia es obvia. Usted entiende las preguntas en castellano porque están expresadas en símbolos cuyos significados le son conocidos. Similarmente, cuando usted da las respuestas en castellano, está produciendo símbolos que son significativos para usted. Pero en el caso del chino no tiene nada de esto. En el caso del chino usted manipula simplemente símbolos formales de acuerdo con un programa de computador y no les añade significado alguno a ninguno de los elementos.

Se han sugerido varias réplicas a este argumento por parte de las personas que trabajan en inteligencia artificial y en psicología, así como en filosofía. Todas ellas

tienen algo en común: todas ellas son inadecuadas. Y hay una razón obvia por la que tienen que ser inadecuadas, ya que el argumento descansa sobre una verdad muy simple, a saber: la sintaxis sola no es suficiente para la semántica y los computadores digitales en tanto que son computadores tienen, por definición, solamente sintaxis.

Quiero clarificar esto considerando un par de argumentos que se presentan a menudo en contra mía.

Algunas personas intentan responder al ejemplo de la habitación china diciendo que la totalidad del sistema entiende chino. La idea es aquí que aunque yo, la persona que está en la habitación manipulando los símbolos, no entiendo chino, yo soy sólo la unidad de procesamiento central del sistema del computador. Ellos argumentan que es todo el sistema, incluyendo la habitación, las cestas llenas de símbolos y los anaqueles que contienen los programas y quizás también otros elementos, tomado como una totalidad, lo que entiende chino. Pero esto está sujeto exactamente a la misma objeción que hice antes. No hay ninguna manera de que el sistema pueda obtener a partir de la sintaxis la semántica. Yo, como unidad de procesamiento central, no tengo ninguna manera de averiguar lo que significa cualquiera de esos símbolos; pero entonces tampoco puede hacerlo todo el sistema.

Otra respuesta común es imaginar que colocamos el programa de comprensión del chino dentro de un robot. Si el robot se moviese e interactuase casualmente con el mundo ¿no sería esto garantía suficiente de que entendía chino? Una vez más la inexorabilidad de la distinción entre semántica y sintaxis derrota esta maniobra. En la medida en que suponemos que el robot tiene solamente un computador por cerebro, aunque pudiese comportarse como si entendiese chino, no habría con todo manera alguna de obtener a partir de la sintaxis la semántica del chino. Esto puede verse si nos imaginamos que yo soy el computador. Dentro de una habitación en el craneo del robot barajo símbolos sin saber que algunos llegan a mí desde cámaras de televisión adosadas a la cabeza del robot y otros salen para mover los brazos y

piernas del robot. En la medida en que todo lo que tengo es un programa de computador formal, no tengo manera de añadirle significado alguno a ninguno de los símbolos. Y el hecho de que el robot esté inmerso en interacciones causales con el mundo exterior no me ayuda a añadir ningún significado a los símbolos a menos que tenga alguna manera de informarse sobre ese hecho. Supongamos que el robot toma una hamburguesa y esto provoca que entre dentro de la habitación el símbolo de una hamburguesa. En la medida en que todo lo que yo tengo es el símbolo sin ningún conocimiento de sus causas o de cómo llegó allí, no tengo ninguna manera de saber lo que significa. Las interacciones causales entre el robot y el resto del mundo son irrelevantes a menos que esas interacciones causales se representen en una mente cualquiera. Pero no hay manera de que puedan serlo si todo en lo que la llamada mente consiste es un conjunto de operaciones puramente formales, sintácticas.

Es importante ver lo que es afirmado y lo que no es afirmado por mi argumento. Supóngase que planteamos la pregunta que mencioné al principio. '¿Puede pensar una máquina?' Bien, en algún sentido, desde luego, todos somos máquinas. Podemos interpretar la materia que tenemos dentro de nuestras cabezas como una máquina de carne. Y desde luego, podemos pensarlo todo. Así, en un sentido de 'máquina', a saber: ese sentido en el que máquina es solamente un sistema físico que es capaz de realizar cierto género de operaciones, en ese sentido, todos somos máquinas, y podemos pensar. Así, trivialmente, hay máquinas que pueden pensar. Pero esta no era la pregunta que nos intrigaba. Así pues, intentemos una formulación diferente de ella. Podría pensar un artefacto? ¿Podría una máquina hecha por el hombre pensar? Bien, una vez más, depende del género de artefacto. Supóngase que hemos diseñado una máquina que fuera molécula-por-molécula indistinguible de un ser humano. Bien, entonces, si se pueden duplicar las causas, entonces presumiblemente pueden duplicarse los efectos. Así, una vez más, la respuesta a esa pregunta es, en principio al menos, trivialmente sí. Si se pudiese construir una máquina que tuviese la misma estructura que un ser

humano, entonces esa máquina sería capaz de pensar. De hecho, sería un substituto de un ser humano. Bien, intentémoslo de nuevo.

La pregunta no es '¿Puede pensar una máquina?' o '¿puede pensar un artefacto?' La pregunta es: '¿Puede pensar un computador digital?' Pero una vez más hemos de ser muy cuidadosos en cómo interpretamos la pregunta. Desde un punto de vista matemático, cualquier cosa puede describirse *como si* fuera un computador digital. Y esto es así porque puede describirse como si instanciase o llevase a cabo un programa de computador. En un sentido completamente trivial, la pluma que está sobre la mesa enfrente de mí puede describirse como un computador digital. Lo único que sucede es que tiene un programa de computador muy aburrido. El programa dice: 'Estáte ahí.' Ahora bien, puesto que en este sentido cualquier cosa es un computador digital, ya que cualquier cosa puede describirse como si estuviera llevando a cabo un programa de computador, entonces, una vez más, nuestra pregunta obtiene una respuesta trivial. Desde luego nuestros cerebros son computadores digitales, puesto que llevan a cabo un número cualquiera de programas de computador. Y, desde luego, nuestros cerebros pueden pensar. Así, una vez más, hay una respuesta trivial a la pregunta. Pero esa no era realmente la pregunta que estábamos intentando plantear. La pregunta que queríamos plantear es ésta: '¿Puede un computador digital, tal como se ha definido, pensar?' Es decir: '¿Es suficiente para, o constitutivo de, pensar el instanciar o llevar a cabo el programa correcto con los *inputs* y *outputs* correctos?' Y a esta pregunta, a diferencia de sus predecesoras, la respuesta es claramente 'no'. Y es 'no' por la razón que hemos puesto de manifiesto reiteradamente, a saber: el programa del computador está definido de manera puramente sintáctica. Pero pensar es algo más que manipular signos carentes de significado, incluye contenidos semánticos significativos. A esos contenidos semánticos es a lo que nos referiremos mediante 'significado'.

Es importante subrayar de nuevo que no estamos hablando sobre un estadio particular del desarrollo de la

tecnología de los computadores. La argumentación no tiene nada que ver con los próximos y pasmosos avances en la ciencia de la computación. No tienen nada que ver con la distinción entre procesos en serie y en paralelo, o con el tamaño de los programas, o la velocidad de las operaciones del computador, o con computadores que pueden interaccionar casualmente con su entorno, o incluso con la invención de robots. El progreso tecnológico se exagera siempre enormemente, pero incluso eliminando la exageración, el desarrollo de los computadores ha sido extraordinariamente notable, y podemos esperar razonablemente que en el futuro se lleven a cabo progresos aún más notables. Sin duda seremos capaces de simular mucho mejor la conducta humana en los computadores de lo que lo podemos hacer en la actualidad, y ciertamente mucho mejor de lo que hemos sido capaces de hacerlo en el pasado. Lo que quiero decir esencialmente es que si estamos hablando sobre tener estados mentales, sobre tener una mente, todas esas simulaciones son simplemente irrelevantes. No importa cuán buena sea la tecnología o cuán rápidos sean los cálculos realizados por el computador. Si se trata realmente de un computador, sus operaciones tienen que definirse sintácticamente, mientras que la conciencia, los sentimientos, los pensamientos, las sensaciones, las emociones, y todo lo demás incluyen algo más que una sintaxis. Por definición el computador es incapaz de *duplicar* esos rasgos por muy poderosa que pueda ser su capacidad para *simular*. La distinción clave aquí es la que se da entre duplicación y simulación. Y ninguna simulación constituye, por sí misma, duplicación.

Lo que he hecho hasta aquí es proporcionar una base al sentido de que esas citas con las que comencé esta charla son realmente tan absurdas como parecen. Hay, sin embargo, una cuestión problemática en esta discusión, y es ésta: '¿Por qué ha pensado alguien alguna vez que los computadores podrían pensar o tener sensaciones y emociones y todo lo demás?' Después de todo, podemos hacer simulaciones computacionales de cualquier proceso del que pueda darse una descripción formal. Así, podemos hacer una simulación computacional

del flujo de dinero en la economía española, o del modelo de distribución de poder en el partido socialista. Podemos hacer una simulación computacional de las tormentas en los términos municipales del país, o de los incendios en los almacenes del este de Madrid. Ahora bien, en cada uno de esos casos, nadie supone que la simulación computacional es efectivamente la cosa real; nadie supone que una simulación computacional de una tormenta nos deje a todos mojados, o que sea probable que una simulación computacional de un incendio vaya a quemar la casa. ¿Por qué diablos va a suponer alguien que esté en sus cabales que una simulación computacional de procesos mentales tiene efectivamente procesos mentales? Realmente desconozco la respuesta a esto, puesto que la idea me parece desde el principio, para decirlo con franqueza, completamente disparatada. Pero puedo hacer un par de especulaciones.

En primer lugar, hay mucha gente que, cuando de la mente se trata, se siente aún tentada a algún tipo de conductismo. Piensan que si algún sistema se comporta como si entendiese chino, entonces realmente tiene que entender chino. Pero ya hemos refutado esta forma de conductismo con el argumento de la habitación china. Otra suposición hecha por mucha gente es que la mente no es parte del mundo biológico, no es parte del mundo de la naturaleza. El punto de vista de la inteligencia artificial fuerte descansa sobre él en su concepción de que la mente es algo puramente formal; que de una manera u otra no puede ser tratada como un producto concreto de procesos biológicos de la misma manera que otro producto biológico. Hay en esas discusiones, para decirlo brevemente, un género de dualismo residual. Los partidarios de IA creen que la mente es algo más que una parte del mundo biológico natural; creen que la mente es especificable de manera puramente formal. La paradoja de esto es que la literatura de IA está llena de recriminaciones contra algún punto de vista llamado 'dualismo', pero de hecho, toda la tesis de la IA fuerte descansa sobre un género de dualismo. Descansa sobre el rechazo de la idea de que la mente

es sólo un fenómeno biológico natural del mundo igual cualquier otro.

Quiero concluir este capítulo uniendo la tesis del capítulo anterior y la tesis de este. Ambas tesis pueden enunciarse de manera muy simple. Y, de hecho, voy a enunciarlas con, quizás, excesiva crudeza. Pero si las unimos pienso que obtenemos una concepción muy poderosa de las relaciones entre mentes, cerebros y computadores. Y el argumento tiene una estructura muy simple, de modo que usted puede ver si es válido o inválido. La primera premisa es:

1. *Los cerebros causan las mentes.*

Ahora bien, esto es realmente demasiado crudo. Lo que queremos decir mediante esto es que los procesos mentales que nosotros consideramos que constituyen una mente son causados, enteramente causados, por procesos que tienen lugar dentro del cerebro. Pero seamos crudos, y abreviemos esto mediante esas cinco palabras —los cerebros causan las mentes. Escribamos ahora la proposición número dos:

2. *La sintaxis no es suficiente para la semántica.*

Esta proposición es una verdad conceptual. Articula justamente nuestra distinción entre la noción de lo que es puramente formal y lo que tiene contenido. Ahora bien, a esas dos proposiciones —que los cerebros causan las mentes y que la sintaxis no es suficiente para la semántica— añadamos una tercera y una cuarta:

3. *Los programas de computador están definidos enteramente por su estructura formal o sintáctica.*

Considero que esta proposición es verdadera por definición, es parte de lo que queremos decir mediante la noción de un programa de computador.

4. *Las mentes tienen contenidos mentales; específicamente, tienen contenidos semánticos.*

Considero que esto es solamente un hecho obvio acerca de cómo funcionan nuestras mentes. Mis pensamientos, y creencias, y deseos son sobre algo, o se refieren a algo, o conciernen a estados de cosas del mundo; y hacen esto porque sus contenidos los dirigen hacia esos estados de cosas del mundo. Ahora bien, a partir de esas cuatro premisas, podemos extraer nuestra primera conclusión; se sigue obviamente de las premisas 2, 3 y 4:

CONCLUSIÓN 1. *Ningún programa de computador es suficiente por sí mismo para dar un sistema, una mente. Los programas, dicho brevemente, no son mentes, y no son suficientes por sí mismos para tener mentes.*

Ahora bien, esto es una conclusión muy poderosa, porque significa que el proyecto de intentar crear mentes diseñando solamente programas está condenado a muerte desde el principio. Y es importante volver a subrayar que esto no tiene nada que ver con ningún estadio particular en el desarrollo de la tecnología, o con ningún estadio particular de la complejidad del programa. Este es un resultado puramente formal, o lógico, obtenido a partir de un conjunto de axiomas en los que están de acuerdo todos (o casi todos) los participantes en la disputa. Es más, incluso la mayor parte de los entusiastas más acérrimos de la inteligencia artifical están de acuerdo en que de hecho, como un asunto de biología, los procesos cerebrales causan estados mentales, y están de acuerdo en que los programas se definen de manera puramente formal. Pero si se unen estas conclusiones con otras cosas que sabemos, entonces se sigue inmediatamente que el proyecto de IA fuerte es incapaz de ser cumplido.

Sin embargo, una vez que hemos obtenido esos axiomas, veamos qué más puede derivarse. He aquí una segunda conclusión:

CONCLUSIÓN 2. *El modo en que las funciones del cerebro causan las mentes no puede ser solamente en virtud de pasar un programa de computador.*

Y en esta segunda conclusión se sigue de poner en conjunción la primera premisa con nuestra primera conclusión. Esto es, del hecho de que los cerebros causan las mentes y del hecho de que los programas no se bastan para llevar a cabo la tarea, se sigue que el modo en que los cerebros causan las mentes no puede ser solamente en virtud de pasar un programa de computador. Ahora bien, pienso que esto es también un resultado importante, puesto que tiene como consecuencia que el cerebro no es, o al menos no es solamente, un computador digital. Vimos anteriormente que cualquier cosa podía describirse trivialmente como si fuera un computador digital, y los cerebros no constituyen ninguna excepción. Pero la importancia de esta conclusión reside en que las propiedades computacionales del cerebro simplemente no bastan para explicar su funcionamiento para producir estados mentales. Y, en efecto, esto debe parecernos una conclusión científica de sentido común en cualquier caso, ya que todo lo que hace es recordarnos el hecho de que los cerebros son motores biológicos; su biología importa. No es solamente un hecho irrelevante sobre la mente, como diversas personas que trabajan en inteligencia artificial han afirmado, el que suceda que está realizada en los cerebros humanos.

Ahora, a partir de nuestra primera premisa, podemos también derivar una tercera conclusión:

CONCLUSIÓN 3. *Cualquier otra cosa que cause las mentes tendría que tener poderes causales equivalentes al menos a los del cerebro.*

Y esta tercera conclusión es una consecuencia trivial de nuestra primera premisa. Es hasta cierto punto similar a decir que si mi motor de gasolina impulsa mi coche a ciento veinte kilómetros/hora, entonces cualquier motor diesel que fuese capaz de hacer eso tendría que

tener una potencia de salida equivalente al menos a la de mi motor de gasolina. Desde luego, algún otro sistema podría causar procesos mentales usando características bioquímicas o químicas enteramente diferentes de las que el cerebro usa de hecho. Podría suceder que hubiese seres en otros planetas, o en otros sistemas solares, que tuviesen estados mentales y usasen una bioquímica enteramente diferente a la nuestra. Supóngase que los marcianos llegasen a la tierra y concluyésemos que tienen estados mentales. Pero supóngase que cuando abriésemos sus cabezas se descubriese que todo lo que había allí dentro era una mucosidad verde. Bien, con todo, la mucosidad verde, si funcionase de manera tal que produjese conciencia y el resto de su vida mental, tendría que tener poderes causales iguales a los del cerebro humano. Ahora bien, de nuestra primera conclusión, que los programas no bastan, y nuestra tercera conclusión, que cualquier otro sistema tendría que tener poderes causales iguales a los del cerebro, se sigue inmediatamente la conclusión cuatro:

CONCLUSIÓN 4. *Para cualquier artefacto que pudiéramos construir que tuviese estados mentales equivalentes a los estados mentales humanos, el desarrollo de un programa de computador no sería suficiente por sí mismo. Más bien, el artefacto tendría que tener poderes equivalentes a los del cerebro humano.*

El resultado de esta discusión es recordarnos algo que ya sabíamos desde el principio, a saber: que los estados mentales son fenómenos biológicos. La conciencia, la intencionalidad y la causación mental son todas ellas parte de la historia de nuestra vida biológica, junto con el crecimiento, la reproducción, la secreción de la bilis y la digestión.

3

LA CIENCIA COGNITIVA

Nos sentimos perfectamente seguros al decir cosas como éstas. 'Basilio votó por los conservadores porque le gustó la manera en que la señora Thatcher llevó el asunto de las islas Malvinas.' Pero no tenemos ni idea de cómo habérnoslas con una situación en la que se dicen 'cosas como ésta: 'Basilio votó por los conservadores a causa de una condición de su hipotálamo.' Esto es, tenemos explicaciones de sentido común de la conducta de las personas en términos mentales, en términos de sus deseos, anhelos, temores, esperanzas y así sucesivamente. Y suponemos que tiene que haber también algún tipo de explicación neuropsicológica de la conducta de las personas en términos de procesos de sus cerebros. El problema es que el primero de esos tipos de explicación funciona bastante bien en la práctica, pero no es científico, mientras que el segundo es ciertamente científico, pero no tenemos ni idea de cómo hacerlo funcionar en la práctica.

Ahora bien, esto nos deja aparentemente con un vacío entre el cerebro y la mente. Y alguno de los mayores esfuerzos intelectuales del siglo XX ha consistido en intentar llenar este vacío, en lograr una ciencia de la conducta humana que no fuera solamente la psicología de sentido común de la abuelita, pero que no fuera tampoco neuropsicología científica. Hasta la época presente, los esfuerzos para llenar el vacío han sido, sin ex-

cepción, fracasos. El conductismo fue el fracaso más espectacular, pero en el transcurso de mi vida he presenciado las pretensiones exageradas hechas a favor de, y eventualmente me he sentido decepcionado por la teoría de juegos, la cibernética, la teoría de la información, el estructuralismo, la sociobiología y un puñado de otras materias. Para anticipar un poco diremos que voy a afirmar que los esfuerzos por llenar el vacío fracasan porque no hay ningún vacío que llenar.

Los esfuerzos más recientes por llenar el vacío descansan sobre analogías entre los seres humanos y los computadores digitales. Según la versión más extrema de este punto de vista, lo que yo llamo 'inteligencia artificial fuerte' o sólo 'IA fuerte', el cerebro es un computador digital y la mente es sólo un programa de computador. Ahora bien, éste es el punto de vista que he refutado en el capítulo anterior. Un intento reciente de llenar el vacío, relacionado con lo anterior, se denomina a menudo 'cognitivismo', puesto que deriva del trabajo en psicología cognitiva e inteligencia artificial, y forma la corriente principal de una nueva disciplina: 'la ciencia cognitiva'. Al igual que la IA fuerte, contempla al computador como la representación correcta de la mente, y no sólo como una metáfora. Pero a diferencia de la IA fuerte, no afirma, o al menos no tiene que afirmar, que los computadores tienen literalmente pensamientos y sensaciones.

Si tuviera que resumirse el programa de investigación del cognitivismo tendría un aspecto parecido a éste: Pensar es procesar información, pero procesar información es solamente manipulación de signos. Los computadores manipulan símbolos. Así, la mejor manera de estudiar el pensar (o, como ellos prefieren llamarlo, 'la cognición') es estudiar los programas computacionales de manipulación de símbolos, ya estén en los computadores o en los cerebros. Según este punto de vista la tarea de la ciencia cognitiva es, entonces, caracterizar el cerebro no al nivel de las células nerviosas ni al nivel de los estados mentales conscientes, sino más bien al nivel de su funcionamiento como un sistema de procesamiento de la información. Y es aquí donde el vacío se rellena.

No puedo exagerar hasta qué punto este proyecto de investigación ha parecido constituir una ruptura importante en la ciencia de la mente. De hecho, de acuerdo con sus defensores, podría ser incluso *la* ruptura que colocase a la psicología en una senda científica segura ahora que se ha liberado de las ilusiones del conductismo.

En esta conferencia voy a atacar al conductismo, pero quiero comenzar por ilustrar su atractivo. Sabemos que hay un nivel de psicología de la abuelita, ingenua, de sentido común y que también hay un nivel de neuropsicología, el nivel de las neuronas, módulos neuronales, sinapsis, neurotransmisores, y todo lo demás. Así pues, ¿por qué tendría alguien que suponer que entre esos dos niveles hay también un nivel de procesos mentales que son procesos computacionales? Y de hecho, ¿por qué habría de suponer alguien que es en ese nivel en el que el cerebro realiza aquellas funciones que consideramos como esenciales para la supervivencia del organismo, a saber: las funciones de procesamiento de la información?

Bien, hay diversas razones: En primer lugar mencionemos una que, en cierto modo, tiene mala fama, pero que pienso que es efectivamente muy influyente. Puesto que no entendemos muy bien el cerebro estamos tentados constantemente a usar la última tecnología como un modelo para intentar entenderlo. En mi niñez se nos aseguraba siempre que el cerebro era una centralita telefónica. ('¿Qué otra cosa podía ser?') Me divertía ver que Sherrington, el gran neurocientífico británico, pensaba que el cerebro funcionaba como un sistema telegráfico. Freud comparaba a menudo el cerebro con los sistemas hidráulicos y electromagnéticos.

Leibniz lo comparaba con un molino, y alguien me dijo que alguno de los antiguos griegos pensaba que el cerebro funcionaba como una catapulta. En la actualidad, obviamente, la metáfora es el computador digital.

Y esto, dicho sea de paso, encaja con las historietas exageradas que, de manera general, oímos hoy en día sobre los computadores y los robots. Se nos asegura frecuentemente en la Prensa sensacionalista que estamos en los umbrales de tener robots domésticos que harán todo

el trabajo del hogar, harán de niñeras para nuestros hijos, nos divertirán con animada conversación y cuidarán de nosotros cuando seamos viejos. Esto, desde luego, tiene bastante de sinsentido. En ninguna parte estamos cerca de ser capaces de producir robots que hagan ninguna de esas cosas. Y de hecho, los robots que han tenido éxito se han limitado a tareas muy restringidas, en contextos muy limitados, tales como las cadenas de producción de automóviles.

Bien, volvamos a las razones serias que tiene la gente para suponer que el cognitivismo es verdadero. En primer lugar, suponen que tienen efectivamente alguna evidencia psicológica de que es verdadero. Hay dos géneros de evidencia. La primera viene de los experimentos de tiempo de reacción; esto es, experimentos que muestran que la gente invierte diferentes sumas de tiempo en realizar tareas intelectuales diferentes. La idea es aquí que si las diferencias en la suma de tiempo que la gente invierte son paralelas a las diferencias de tiempo que invertiría un computador, entonces esto es, al menos, prueba de que el sistema humano está trabajando sobre los mismos principios que un computador. El segundo tipo de prueba viene de la lingüística, especialmente del trabajo de Chomsky y sus colegas sobre gramática generativa. La idea es aquí que las reglas formales de gramática que la gente sigue cuando habla un lenguaje son semejantes a las reglas formales que sigue un computador.

No voy a decir muchas cosas sobre la prueba del tiempo de reacción, puesto que pienso que todo el mundo está de acuerdo en que es completamente no concluyente y que está sujeta a una gran cantidad de interpretaciones diferentes. Diré algo sobre la prueba lingüística.

Sin embargo, de manera subyacente a la interpretación computacional de ambos géneros de prueba está una razón mucho más profunda, y creo que más influyente para aceptar el cognitivismo. La segunda razón es una tesis general que se supone que ejemplifican los dos géneros de prueba, y que reza como sigue: puesto que podemos diseñar computadores que siguen reglas cuando procesan información, y puesto que aparentemente los

seres humanos siguen también reglas cuando piensan, entonces hay algún sentido unitario en el que el cerebro y el computador están funcionando de una manera similar e incluso puede ser que la misma.

La tercera suposición que está detrás del programa de investigación cognitivista es una vieja suposición. Se remonta hasta Leibniz y quizá hasta Platón. Es la suposición de que un logro mental tiene que tener causas teóricas. Es la suposición de que si el resultado de un sistema es *significativo,* en el sentido de que, por ejemplo, nuestra capacidad para aprender un lenguaje o nuestra capacidad para reconocer rostros es una capacidad cognitiva significativa, entonces tiene que haber alguna teoría, internalizada de alguna manera en nuestros cerebros, que subyace a esa actividad.

Finalmente, hay otra razón por la que la gente se adhiere al programa de investigación cognitivista, especialmente si tienen inclinación filosófica. No pueden ver otra manera de entender las relaciones entre la mente y el cerebro. Puesto que entendemos la relación de un programa del computador con el *hardware* del computador, esto nos proporciona un modelo excelente, puede ser que el único modelo, que nos capacitará para explicar las relaciones entre la mente y el cerebro. Ya he respondido a esta pretensión en el primer capítulo, de modo que no necesito discutirla más aquí.

Bien, ¿qué haremos con todos esos argumentos a favor del cognitivismo? No creo que tenga una refutación abrumadora del cognitivismo en el sentido que creo que tengo una de la IA fuerte. Pero creo que si examinamos los argumentos que se dan a favor del cognitivismo, veremos que son muy débiles. Y, efectivamente, un desenmascaramiento de su debilidad nos capacitará para entender varias diferencias importantes entre el modo en que los seres humanos se comportan y el modo en que funcionan los computadores.

Comencemos con la noción de seguir una regla. Se nos dice que los seres humanos siguen reglas y que los computadores siguen reglas. Pero quiero argumentar que hay una diferencia crucial. En el caso de los seres humanos, siempre que seguimos una regla, estamos siendo

guiados por el contenido efectivo o el significado de la regla. En el caso humano de seguir una regla, los significados causan la conducta. Ahora bien, desde luego no causan la conducta enteramente por sí misma, pero ciertamente juegan un papel causal en la producción de la conducta. Considérese, por ejemplo, la regla: conduce por la izquierda de la carretera en Gran Bretaña. Ahora bien, siempre que vengo a Gran Bretaña tengo que recordarme a mí mismo esta regla. ¿Cómo funciona? Decir que estoy obedeciendo la regla es decir que el significado de la regla, esto es, su contenido semántico, juega algún género de papel causal en la producción de lo que efectivamente hago. Obsérvese que hay montones de otras reglas que describirían lo que está sucediendo. Pero no son reglas tales que suceda que las estoy siguiendo. Así, por ejemplo, suponiendo que estoy en una carretera con dos carriles y que el volante está colocado en el lado derecho del coche, entonces podría decirse que mi conducta está de acuerdo con la regla: conduce de tal manera que el volante esté lo más cerca posible a la línea central de la carretera. Ahora bien, esto es de hecho una descripción correcta de mi conducta. Pero esta no es la regla que estoy siguiendo en Gran Bretaña. La regla que sigo es: conduce por el lado izquierdo de la carretera.

Quiero que este punto esté completamente claro, así que permítaseme dar otro ejemplo. Cuando mis hijos fueron a la academia de conducir de Oakland, se les enseñó una regla para aparcar coches. La regla era: haz maniobrar tu coche hacia el bordillo con el volante girado completamente hacia la derecha hasta que tus ruedas delanteras estén en línea con las ruedas traseras del coche que está enfrente de ti. A continuación, gira completamente el volante hacia la izquierda. Ahora bien, obsérvese que si están siguiendo esa regla, entonces su significado debe jugar algún papel causal en la producción de sus conductas. Yo estaba interesado en aprender esta regla porque se trata de una regla que yo no sigo. De hecho, yo no sigo ninguna regla en absoluto cuando aparco el coche. Solamente miro al bordillo e intento llegar tan cerca de él como pueda sin causar

destrozos en los coches que están enfrente y detrás de mí. Pero obsérvese que podría suceder que mi conducta, vista desde fuera, vista externamente, fuese idéntica a la conducta de la persona que está siguiendo la regla. Con todo, no sería verdadero decir de mí que yo estaba siguiendo la regla. Las propiedades formales de la conducta no son suficientes para mostrar que se está siguiendo una regla. Para que la regla se siga, el significado de la regla ha de jugar algún papel causal en la conducta.

Ahora bien, la moraleja de esta discusión sobre el cognitivismo puede expresarse de manera muy simple: *En el sentido en que los seres humanos siguen reglas* (y, a propósito, los seres humanos siguen reglas en mucha menor medida de la que los cognitivistas dicen que las siguen), *en este sentido los computadores no siguen reglas en absoluto. Solamente actúan de acuerdo con ciertos procedimientos formales.* El programa del computador determina los diversos pasos que debe seguir la maquinaria; determina cómo se transformará un estado en un estado subsiguiente. Y podemos hablar *metafóricamente* como si fuera un asunto de seguir reglas. Pero en el sentido *literal* en el que los seres humanos siguen reglas los computadores no siguen reglas, actúan solamente como si estuviesen siguiendo reglas. Ahora bien, tales metáforas son completamente inocuas y son a la vez comunes y útiles en ciencia. Podemos hablar metafóricamente de un sistema, del sistema solar, por ejemplo, como si estuviera siguiendo reglas. La metáfora sólo se convierte en perjudicial cuando se confunde con el sentido literal. Está muy bien usar una metáfora psicológica para explicar el computador. La confusión viene cuando se toma la metáfora literalmente y se usa el sentido metafórico de seguir una regla por parte del computador para intentar explicar el sentido psicológico de seguir una regla, sobre el que se basaba la metáfora en primer lugar.

Estamos ahora en posición de decir qué es lo que estaba equivocado en la evidencia lingüística a favor del cognitivismo. Si es de hecho cierto que la gente sigue reglas de sintaxis cuando habla, esto no muestra que se

comporte como los computadores digitales, puesto que en el sentido en que la gente sigue reglas de sintaxis, el computador no las sigue en absoluto. El solamente pasa a través de procesos formales.

Así pues, tenemos dos sentidos de seguir una regla, uno literal y otro metafórico. Y es muy fácil confundir los dos. Ahora bien, quiero aplicar estas lecciones a la noción de procesamiento de información. Creo que la noción de procesamiento de información incorpora una confusión de bulto similar. La idea es que puesto que yo proceso información cuando pienso y puesto que mi máquina de calcular procesa información cuando considera algo como *input,* lo transforma y produce información como *output,* entonces tiene que haber algún sentido unitario en el que ambos estamos procesando información. Pero esto me parece obviamente falso. El sentido en el que yo hago procesamiento de información cuando pienso es el sentido en el que estoy consciente o inconscientemente ocupado en cierto proceso mental. Pero en este sentido de procesamiento de información, la calculadora no hace procesamiento de información, puesto que no tiene en absoluto ningún proceso mental. Simplemente remeda o simula los rasgos formales del proceso mental que yo tengo. Esto es, incluso si los pasos por los que atraviesa la calculadora son formalmente los mismos pasos que yo atravieso, esto no mostraría que la máquina hace algo parecido en absoluto a lo que yo hago, por la simple razón de que la calculadora no tiene fenómenos mentales. Al sumar 6 y 3 ella no sabe que el numeral '3' está por el número tres, y que el signo más está por la operación de adición. Y esto es así por la muy simple razón de que ella no sabe nada. Es más, a esto se debe el que tengamos calculadoras. Pueden hacer los cálculos más rápidos y más exactamente que nosotros sin tener que pasar por ningún esfuerzo mental para hacerlo. En el sentido en que tenemos que pasar por el procesamiento de información, ellas no tienen que pasar.

Necesitamos, entonces, hacer una distinción entre dos sentidos de la noción de procesamiento de información. O al menos, dos géneros radicalmente distintos de pro-

cesamiento de información. El primer género, que llamaré 'procesamiento de información psicológica', incluye estados mentales. Para decirlo de la manera más cruda: cuando la gente realiza operaciones mentales, piensa efectivamente, y pensar incluye característicamente procesamiento de información de un género u otro. Pero hay otro sentido de procesamiento de información en el que no hay estados mentales en absoluto. En esos casos, hay procesos que son *como si* ocurriese algún procesamiento mental de información. Llamemos a este segundo género de casos de procesamiento de información formas de procesamiento de información 'como si'. Es perfectamente inocuo usar cualesquiera de esos dos géneros de adscripciones mentales en el supuesto de que no las confundamos. Sin embargo, lo que encontramos en el cognitivismo es una persistente confusión entre los dos.

Ahora bien, una vez que vemos esta distinción claramente vemos una de las más profundas debilidades del argumento cognitivista. Del hecho de que yo proceso información cuando pienso, y del hecho de que el computador procesa información —incluso procesamiento de información que puede simular los rasgos formales de mi pensar— simplemente no se sigue que haya nada psicológicamente relevante sobre el programa del computador. Para mostrar la relevancia psicológica tendría que haber algún argumento independiente acerca de que el procesamiento de información computacional es psicológicamente relevante. La noción de procesamiento de información se usa para enmascarar esta confusión, puesto que una expresión está siendo usada para cubrir dos fenómenos completamente distintos. Dicho brevemente, la confusión que encontramos sobre seguir una regla tiene un paralelo exacto en la noción de procesamiento de información.

Sin embargo, hay una confusión más profunda y más sutil incluida en la noción de procesamiento de información. Obsérvese que en el sentido de procesamiento de información 'como si', un sistema cualquiera puede describirse como si estuviera haciendo procesamiento de información, y de hecho, podríamos usarlo incluso para reunir información. Así pues, no es solamente un asunto

de usar calculadoras y computadores. Considérese, por ejemplo, el agua corriente cuesta abajo. Ahora bien, podemos describir el agua como si estuviese haciendo procesamiento de información. Y podríamos incluso usarla para obtener información. Podríamos usarla, por ejemplo, para obtener información sobre la línea de menor resistencia de los contornos de la colina. Pero no se sigue de esto que el agua que corre cuesta abajo tenga alguna relevancia psicológica. No hay psicología alguna en la acción de la gravedad sobre el agua.

Pero podemos aplicar las lecciones de este punto al estudio del cerebro. Es un hecho obvio que el cerebro tiene un nivel de procesos de información psicológica real. Para repetirlo: la gente piensa efectivamente y el pensar se desarrolla en los cerebros de las personas. Además, hay todo tipo de cosas que se desarrollan en el cerebro en el nivel neurofisiológico que causan efectivamente nuestros procesos de pensamiento. Pero mucha gente supone que además de esos dos niveles, el nivel de la psicología ingenua y el nivel de la neurofisiología tiene que haber algún nivel adicional de procesamiento de información computacional. Ahora bien, ¿por qué suponen esto? Creo que en parte porque confunden el nivel psicológicamente real de procesamiento de la información con la posibilidad de dar descripciones de procesamiento de información 'como si' de los procesos que se desarrollan en el cerebro. Si se habla de agua corriendo cuesta abajo, todo el mundo puede ver que esto es psicológicamente irrelevante. Pero es difícil ver que el mismo punto se aplica exactamente al cerebro.

Lo que es psicológicamente relevante sobre el cerebro es el hecho de que contiene procesos psicológicos y el hecho de que tiene una neurofisiología que causa y realiza esos procesos. Pero el hecho de que podamos describir otros proceso en el cerebro desde un punto de vista de procesamiento de información 'como si', no proporciona por sí mismo ninguna prueba de que esos procesos son psicológicamente reales o incluso psicológicamente relevantes. Una vez que estamos hablando sobre el interior del cerebro, es difícil ver la confusión, pero se trata de exactamente de la misma confusión que la

confusión de suponer que puesto que el agua que corre cuesta abajo hace procesamiento de información 'como si', hay alguna psicología oculta en el hecho de que el agua corra cuesta abajo.

La siguiente suposición a examinar es la idea de que detrás de toda conducta significativa tiene que haber alguna teoría interna. Esta suposición se encuentra en muchas áreas y no solamente en psicología cognitiva. Así, por ejemplo, la búsqueda por parte de Chomsky de una gramática universal está basada sobre la suposición de que si hay ciertos rasgos comunes a todos los lenguajes y si esos rasgos están constreñidos por rasgos comunes del cerebro humano, entonces tiene que haber en el cerebro un entero y complejo conjunto de reglas de gramática universal. Pero una hipótesis mucho más simple sería que la estructura fisiológica del cerebro impone las gramáticas posibles sin la intervención de un nivel intermedio de reglas o teorías. Esta hipótesis no solamente es más simple, sino que también la misma existencia de rasgos universales de lenguaje constreñidos por rasgos innatos del cerebro sugiere que el nivel de descripción neurofisiológico es suficiente. No se necesita suponer que hay regla alguna encima de las estructuras neurofisiológicas.

Espero que un par de analogías aclarará esto. Constituye un hecho simple sobre la visión humana el que no podamos ver infrarrojos o ultravioletas. Ahora bien, ¿sucede esto así porque tenemos una regla universal de gramática visual que dice: 'No ver infrarrojos o ultravioletas'? No, obviamente esto sucede así porque nuestro aparato visual simplemente no es sensible a esos dos extremos del espectro. Desde luego, podríamos describirnos a nosotros mismos *como si* estuviésemos siguiendo una regla de gramática visual, pero a pesar de todo, no lo estamos. O, para tomar otro ejemplo, si intentásemos hacer un análisis teórico de la capacidad humana para guardar el equilibrio mientras se está andando, parecería como si allí se estuvieran desarrollando algunos procesos mentales más o menos complejos, como si tomando en cuenta claves de varios géneros resolviésemos una serie de ecuaciones de segundo grado, inconscientemente desde luego, y eso nos capacitase para andar sin caernos. Pero de

hecho sabemos que este tipo de teoría mental no es necesaria para dar cuenta del logro de andar sin caernos. De hecho, esto se lleva a cabo en gran parte por unos fluidos que están situados en el oído interno y que simplemente no calculan nada en absoluto. Si uno da vueltas lo suficiente de modo que los fluidos se perturben, es probable que caiga. Ahora bien, quiero sugerir que gran parte de nuestros logros cognitivos pueden muy bien ser semejantes a éste. El cerebro, ni más ni menos, los hace. No tenemos ninguna buena razón para suponer que además del nivel de nuestros estados mentales y del nivel de nuestra neurofisiología se está desarrollando algún cálculo inconsciente.

Considérese el reconocimiento de rostros. Todos nosotros reconocemos los rostros de nuestros amigos, parientes y conocidos sin esfuerzo alguno de ningún tipo, y de hecho ahora tenemos pruebas de que ciertas porciones del cerebro están especializadas en reconocimiento de rostros. ¿Cómo funciona esto? Bien, supóngase que vamos a diseñar un computador que pudiese reconocer rostros tal como nosotros lo hacemos. Ello comportaría una tarea computacional completa, incluyendo un gran número de cálculos de rasgos geométricos y topológicos. Pero ¿constituye esto una evidencia de que el modo en que nosotros lo hacemos incluye cálculo y computación? Obsérvese que cuando pisamos en arena húmeda y dejamos una huella, ni nuestro pie ni la arena hacen computación alguna. Pero si fuésemos a diseñar un programa que calculase la topología de una huella a partir de la información sobre presiones diferenciales sobre la arena, esto constituiría una tarea computacional bastante compleja. El hecho de que una simulación computacional de un fenómeno natural incluya un procesamiento de información complejo no muestra que el fenómeno mismo incluya tal procesamiento. Y puede ser que el reconocimiento de rostros sea tan simple y tan automático como el dejar huellas en la arena.

De hecho, si proseguimos consecuentemente la analogía del computador, encontraremos que hay muchísimas cosas que se desarrollan en el computador que no son procesos computacionales tampoco. Por ejemplo, en el caso

de algunas calculadoras, si se pregunta: '¿Cómo multiplica la calculadora siete por tres?', la respuesta es: 'Añade tres a sí mismo siete veces.' Pero si se pregunta entonces: '¿Y cómo añade tres a sí mismo?', no hay respuesta computacional alguna a esto; simplemente está hecho en el *hardware*. Así la respuesta a esta pregunta es: 'Pura y simplemente, lo hace.' Y quiero sugerir que para muchísimas capacidades absolutamente fundamentales, tales como nuestra capacidad para ver o nuestra capacidad para aprender un lenguaje, no puede haber ningún nivel mental teórico subyacente a esas capacidades: el cerebro, pura y simplemente, las hace. Estamos neurofisiológicamente construidos de tal manera que el asalto de los fotones sobre nuestras células fotorreceptoras nos capacita para ver, y estamos neurofisiológicamente construidos de tal manera que la estimulación procedente de oír a otra gente hablar y de interaccionar con ellos nos capacita para aprender un lenguaje.

Ahora bien, no estoy diciendo que las reglas no jueguen ningún papel en nuestra conducta. Por el contrario, las reglas de lenguaje o reglas de juegos, por ejemplo, parecen jugar un papel crucial en la conducta relevante. Pero estoy diciendo que es una cuestión delicada el decidir qué partes de la conducta están gobernadas por reglas y cuáles no. Y no podemos suponer que toda conducta significativa tenga, de manera suyacente a ella, algún sistema de reglas.

Quizá éste es un buen lugar para decir que aunque no soy optimista sobre el proyecto de investigación global del cognitivismo, pienso que es probable que una buena porción de intuiciones fruto del esfuerzo serán aprovechables, y ciertamente no quiero desanimar a nadie de que intente probar que estoy equivocado. E incluso si estoy en lo cierto, muchas intuiciones procedentes de proyectos de investigación fallidos pueden aprovecharse; el conductismo y la psicología freudiana son dos casos a tener en cuenta. Me ha impresionado especialmente la obra de David Marr sobre la visión y la obra de varias personas sobre 'comprensión del lenguaje natural', esto es, sobre el esfuerzo para lograr que los computadores

simulen la producción y la interpretación del habla humana ordinaria.

Quiero concluir este capítulo con una nota más positiva diciendo cuáles son las implicaciones de este enfoque para el estudio de la mente. Como un modo de contrarrestar el cuadro cognitivista, permítaseme presentar un enfoque alternativo a la solución de los problemas que acosan a las ciencias sociales. Abandonemos la idea de que hay un programa de computador entre la mente y el cerebro. Pensemos en la mente y los procesos mentales como fenómenos biológicos que están biológicamente basados como el crecimiento, la digestión o la secreción de la bilis. Pensemos en nuestra experiencia visual, por ejemplo, como el producto final de una serie de eventos que comienzan con el asalto de los fotones a la retina y terminan en algún lugar del cerebro. Ahora bien, habrá dos grandes niveles de descripción en la explicación causal de cómo tiene lugar la visión en los animales. Habrá un primer nivel de la neurofisiología; un nivel en el que podemos discutir sobre neuronas individuales, sinapsis y potenciales de acción. Pero dentro de este nivel neurofisiológico habrá niveles inferiores y superiores de descripción. No es necesario limitarnos solamente a las neuronas y a las sinapsis. Podemos hablar sobre la conducta de grupos o módulos de neuronas, tales como los diferentes niveles de tipos de neuronas de la retina o las columnas del córtex, y podemos hablar sobre la realización de los sistemas neurofisiológicos a niveles mucho mayores de complejidad, tales como el papel de córtex estriado en la visión, o el papel de las zonas 18 y 19 del córtex visual, o las relaciones entre el córtex visual y el resto del cerebro al procesar estímulos visuales. Así, dentro del nivel neurofisiológico habrá una serie de niveles de descripción, todos ellos igualmente neurofisiológicos.

Ahora bien, además de esto habrá también un nivel mental de descripción. Sabemos, por ejemplo, que la percepción es una función de expectatividad. Si se espera ver algo, se verá mucho más fácilmente. Sabemos además que la percepción puede estar afectada por varios fenómenos mentales. Sabemos que el humor y la emoción

pueden afectar a cómo y qué se percibe. Y de nuevo, dentro de este nivel mental, habrá diferentes niveles de descripción. Podemos hablar no sólo sobre cómo es afectada la percepción por creencias individuales y deseos, sino también sobre cómo es afectada por fenómenos mentales globales, tales como el *background* de capacidades de la persona o su perspectiva general sobre el mundo. Pero además del nivel de la neurofisiología, y del nivel de la intencionalidad, no necesitamos suponer que haya otro nivel, un nivel del proceso computacional digital. Y no hay daño alguno en pensar tanto en el nivel de los estados mentales como en el nivel de la neurofisiología como procesamiento de información, siempre que no cometamos la confusión de suponer que la forma psicológica real del procesamiento de información es la misma que la 'como si'.

Para concluir, ¿dónde estamos en nuestra valoración del programa de investigación cognitivista? Bien, ciertamente no he demostrado que sea falso. Podría suceder que fuese verdadero. Pienso que sus oportunidades de éxito son casi tan grandes como las oportunidades de éxito del conductismo hace quince años. Es decir, pienso que sus oportunidades de éxito son prácticamente cero. Con todo, lo que he hecho al argumentar a favor de esto es simplemente plantear las tres cosas siguientes: en primer lugar he sugerido que una vez que se ponen al descubierto las suposiciones básicas que están detrás del cognitivismo, su implausibilidad es completamente aparente. Pero las suposiciones están, en gran parte, muy profundamente asentadas en nuestra cultura intelectual, algunas de ellas son muy difíciles de desarraigar o incluso de ser completamente consciente de ellas. Mi primera afirmación es que una vez que entendemos completamente la naturaleza de las suposiciones, su implausibilidad resulta manifiesta. La segunda cuestión en la que he insistido es que de hecho no tenemos suficiente evidencia empírica para suponer que esas suposiciones son verdaderas. Puesto que la interpretación de la evidencia existente descansa sobre una ambigüedad en ciertas nociones cruciales, tales como las de procesamiento de información y de seguir una regla. Y en tercer lugar, he presen-

tado un punto de vista alternativo, tanto en este capítulo como en el primero, de las relaciones entre el cerebro y la mente; un punto de vista que no exige postular ningún nivel intermedio de proceso computacional algorítmico que medie entre la neurofisiología del cerebro y la intencionalidad de la mente. La característica de este cuadro que es importante para esta discusión, es que además de un nivel de estados mentales, tales como creencias y deseos, y un nivel de neurofisiología, no hay ningún otro nivel, no se necesita nada que rellene el hueco entre la mente y el cerebro, porque no hay hueco que rellenar. Como metáfora para el cerebro el computador no es probablemente ni mejor ni peor que anteriores metáforas mecánicas. Aprendemos tanto sobre el cerebro diciendo que es un computador como diciendo que es una centralita telefónica, un sistema telegráfico, una bomba de agua o un motor de vapor.

Supóngase que nadie supiese cómo funcionan los relojes. Supóngase que fuese terriblemente difícil averiguar cómo funcionan, porque, aunque hubiese muchos por todas partes, nadie supiera cómo hacer uno, y los esfuerzos por averiguar cómo funcionan tendieran a destruir el reloj. Supóngase ahora que un grupo de investigadores dijese: 'Entenderemos cómo funcionan los relojes si diseñamos una máquina que sea el equivalente funcional de un reloj, que señale el tiempo tan bien como un reloj.' De este modo, diseñaron un reloj de arena y afirmaron: 'Ahora entendemos cómo funcionan los relojes', o quizá: '¡Ojalá pudiésemos lograr que el reloj de arena fuese tan exacto como un reloj!; en ese caso entenderíamos, al fin, cómo funcionan los relojes.' Substitúyase 'reloj' por 'cerebro' en esta parábola, y substitúyase 'reloj de arena' por 'computador digital' y la noción de inteligencia por la de señalar el tiempo y tendremos en gran parte (¡no totalmente!) la situación contemporánea de la inteligencia artificial y de la ciencia cognitiva.

Mi objetivo global en esta investigación es intentar responder alguna de las cuestiones más problemáticas sobre cómo los seres humanos encajan en el resto del universo. En el primer capítulo intenté resolver el 'problema mente-cuerpo'. En el segundo me deshice de algunas

pretensiones extremas que identifican a los seres humanos con los computadores digitales. En éste he planteado algunas dudas sobre el programa de investigación cognitivista. En la segunda parte del libro quiero dirigir mi atención a explicar la estructura de las acciones humanas, la naturaleza de las ciencias sociales y los problemas del libre albedrío.

4

LA ESTRUCTURA DE LA ACCIÓN

El propósito de este capítulo es explicar la estructura de la acción humana. Necesito hacer esto por varias razones. Necesito mostrar cómo la naturaleza de la acción es consecuente con mi explicación del problema mente-cuerpo y con mi rechazo de la inteligencia artificial, contenido en los capítulos anteriores. Necesito explicar el componente mental de la acción y mostrar cómo se relaciona con el componente físico. Necesito mostrar cómo se relaciona la estructura de la acción con la explicación de la acción. Y necesito establecer un fundamento para la discusión de la naturaleza de las ciencias sociales y la posibilidad del libre albedrío, que discutiré en los dos últimos capítulos.

Si pensamos sobre las acciones humanas, encontramos inmediatamente algunas diferencias chocantes entre ellas y otros eventos del mundo natural. A primera vista resulta tentador pensar qué tipos de acciones o de conducta pueden identificarse con tipos de movimientos corporales. Pero esto es obviamente erróneo. Por ejemplo, uno y el mismo conjunto de movimientos corporales humanos podría constituir una danza, o la realización de señales, o el hacer ejercicio, o el probar los propios músculos, o nada de lo anterior. Además, lo mismo que uno y el mismo tipo de movimientos físicos pueden constituir géneros completamente diferentes de acciones, así también un tipo de acción puede ser realizada por un

número sumamente diferente de tipos de movimientos físicos. Piénsese, por ejemplo, en el de enviar un mensaje a un amigo. Se puede escribir en una hoja de papel. Se puede mecanografiar. Puede enviarse por un mensajero o mediante un telegrama. O se puede dar de viva voz a través del teléfono. Y de hecho, cada uno de esos medios de enviar el mismo mensaje podría efectuarse con una variedad de movimientos físicos. Uno podría escribir la nota con su mano izquierda o con su mano derecha, con los dedos del pie o incluso sujetando el lápiz entre sus dientes. Además, otra característica extraña de las acciones que las hace diferentes de los eventos en general es que las acciones parecen tener descripciones preferentes. Si estoy dando un paseo hacia el Retiro, hay un cierto número de otras cosas que están sucediendo en el curso de mi paseo, pero sus descripciones no describen mis acciones intencionales, puesto que al actuar lo que estoy haciendo depende en gran parte de lo que pienso que estoy haciendo. Así, por ejemplo, me estoy desplazando también en la dirección general de la Patagonia, sacudiendo hacia arriba y hacia abajo el cabello de mi cabeza, desgastando las suelas de mis zapatos y moviendo un gran cantidad de moléculas de aire. Sin embargo, no me parece que ninguna de esas otras descripciones vaya a lo que es esencial en esta acción, tal como la acción es.

Una tercera característica de las acciones, relacionada con las anteriores, es que una persona está en una posición especial para saber lo que está haciendo. No necesita observarse a sí misma o llevar a cabo ninguna investigación para ver qué acción está realizando, o al menos está intentando realizar. Si se me dice: '¿Estás intentando dar un paseo hacia el Retiro o estás intentando acercarte a la Patagonia?', no tengo duda alguna al dar una respuesta, aunque los movimientos físicos que hago podrían ser apropiados para cualquiera de las dos respuestas.

Es también un hecho destacable de los seres humanos el que sin esfuerzo alguno de ningún tipo seamos capaces de identificar y explicar nuestra conducta y la de otras personas. Creo que esta capacidad descansa en nues-

tro dominio inconsciente de un cierto conjunto de principios, lo mismo que nuestra capacidad de reconocer algo como una oración del castellano descansa sobre el hecho de que tenemos un dominio inconsciente de los principios de la gramática castellana. Creo que hay un conjunto de principios que presuponemos cuando decimos cosas de sentido común ordinario, tales como, por ejemplo, que Basilio votó por los conservadores porque pensaba que ellos atajarían el problema de la inflación, o que María se trasladó de Bilbao a Barcelona porque pensaba que las oportunidades de trabajo eran allí mejores, o incluso cosas tan simples como que ese hombre de allí que está haciendo esos movimientos tan extraños está, de hecho, afilando un hacha, o sacando brillo a sus zapatos.

Es común que la gente que reconoce la existencia de esos principios teóricos se burle de ellos diciendo que son solamente una teoría folklórica y que deberían ser sustituidos por alguna explicación más científica de la conducta humana. Esta afirmación me resulta sospechosa, lo mismo que me resultaría sospechosa una afirmación que dijese que deberíamos sustituir nuestra teoría implícita de la gramática del castellano, la que adquirimos al aprender ese lenguaje. La razón de mi sospecha es la misma en cada caso: usar la teoría implícita forma parte de realizar la acción, del mismo modo que usar las reglas de gramática es parte de lo que se hace al hablar. Así, aunque podríamos añadirle a ella, o descubrir, todo tipo de cosas adicionales interesantes sobre el lenguaje o sobre la conducta, es muy poco probable que podamos reemplazar esa teoría que está implícita, y es parcialmente constitutiva del fenómeno, por alguna explicación 'científica' externa de ese mismo fenómeno.

Aristóteles y Descartes habrían estado completamente familiarizados con la mayor parte de nuestras explicaciones de la conducta humana, pero no con nuestras explicaciones de los fenómenos biológicos y físicos. La razón usualmente aducida para esto es que tanto Aristóteles como Descartes tenían una teoría primitiva de la biología y de la física por un lado, y una teoría primitiva de la conducta humana por el otro; y mientras que nosotros hemos avanzado en biología y en fí-

sica, no hemos hecho avances comparables en la explicación de la conducta humana. Quiero sugerir un punto de vista alternativo. Pienso que Aristóteles y Descartes, al igual que nosotros mismos, tenían ya una teoría compleja y sofisticada de la conducta humana. Pienso también que muchas explicaciones supuestamente científicas de la conducta humana, tales como las de Freud, de hecho emplean más bien que reemplazan los principios de nuestra teoría implícita de la conducta humana.

Resumamos lo que he dicho hasta ahora: Hay más tipos de acción que tipos de movimientos físicos, las acciones tienen descripciones preferenciales, la gente sabe lo que está haciendo sin necesidad de observación, y los principios mediante los cuales identificamos y explicamos la acción son ellos mismos parte de las acciones; esto es, son parcialmente constitutivos de las acciones. Quiero ahora dar una breve explicación de lo que podríamos llamar la estructura de la conducta.

Para explicar la estructura de la conducta humana, necesito introducir uno o dos términos técnicos. La noción clave en la estructura de la conducta es la noción de intencionalidad. Decir que un estado mental tiene intencionalidad significa simplemente que es sobre algo. Por ejemplo, una creencia es siempre una creencia de que tal y tal es el caso, o un deseo es siempre un deseo de que tal y tal suceda o sea el caso. Intentar, en el sentido ordinario, no tiene ningún papel especial en la teoría de la intencionalidad. Intentar hacer algo es solamente un género de intencionalidad junto con creer, desear, esperar, temer y así sucesivamente.

Un estado intencional como una creencia, o un deseo, o una intención en el sentido ordinario, tiene característicamente dos componentes. Tiene lo que podríamos llamar su contenido, que lo hace ser sobre algo, y su 'modo psicológico' o 'tipo'. La razón por la que necesitamos esta distinción es que se puede tener el mismo contenido en tipos diferentes. Así, por ejemplo, puedo querer salir de la habitación, puedo creer que saldré de la habitación, puedo intentar salir de la habitación. En cada caso tenemos el mismo contenido, que saldré de la

habitación, pero en diferentes modos o tipos psicológicos: creencia, deseo e intención, respectivamente.

Además, el contenido y el tipo del estado servirá para relacionar el estado mental con el mundo. Después de todo a esto se debe que tengamos mentes con estados mentales: para representarnos el mundo a nosotros mismos; para representar cómo es, cómo quisiéramos que fuese, en qué tememos que pueda convertirse, lo que intentamos hacer respecto de él y así sucesivamente. Esto tiene como consecuencia que nuestras creencias serán verdaderas si se acoplan con el modo en que es el mundo, falsas si no lo hacen; nuestros deseos serán cumplidos o frustrados, nuestras intenciones llevadas a cabo o no llevadas a cabo. En general, los estados intencionales tienen, entonces, 'condiciones de satisfacción'. Cada estado mismo determina bajo qué condiciones es verdadero (si, pongo por caso, se trata de una creencia) o bajo qué condiciones es llevado a cabo (si se trata de una intención). En cada caso el estado mental representa sus propias condiciones de satisfacción.

Una tercera característica que ha de observarse sobre tales estados es que algunas veces causan que sucedan cosas. Por ejemplo, si quiero ir al cine, y de hecho voy al cine, normalmente mi deseo causará el mismo evento que él representa, mi ir al cine. En tales casos hay una conexión interna entre la causa y el efecto, puesto que la causa es una representación del mismo estado de cosas que ella causa. La causa representa y a la vez ocasiona el efecto. Llamo a tales géneros de relaciones de causa y efecto, caso de 'causación intencional'. La causación intencional resultará crucial, como veremos, tanto para la estructura como para la explicación de la acción humana. Es, en varios sentidos, completamente diferente de las explicaciones estándar de los libros de texto de la causación, donde, por ejemplo, una bola de billar golpea a otra bola de billar y causa que esta última se mueva. Para nuestros propósitos lo esencial sobre la causación intencional es que en los casos que consideraremos la mente ocasiona el mismo estado de cosas sobre el que ha estado pensando.

Resumamos esta discusión de la intencionalidad: hay

tres características que necesitamos tener presentes en nuestro análisis de la conducta humana. En primer lugar, los estados intencionales constan de un contenido en un cierto tipo mental. En segundo lugar, ellos determinan sus condiciones de satisfacción; esto es, serán satisfechos o no dependiendo de si el mundo se acopla con el contenido del estado. Y en tercer lugar, algunas veces ellos causan que sucedan cosas, por medio de la causación intencional para dar lugar a un acople; esto es, para dar lugar al estado de cosas que representan, a sus propias condiciones de satisfacción.

Usando esas ideas volveré ahora a la principal tarea de este capítulo. He prometido dar una explicación muy breve de lo que podría llamarse la estructura de la acción, o de la estructura de la conducta. Por conducta entiendo aquí conducta humana, intencional, voluntaria. Me refiero a cosas tales como pasear, correr, comer, hacer el amor, votar en las elecciones, casarse, comprar y vender, ir de vacaciones, trabajar en un empleo. No me refiero a cosas tales como hacer la digestión, envejecer o roncar. Pero incluso si nos restringimos a la conducta intencional, las actividades humanas se nos presentan con una desconcertante variedad de tipos. Necesitaremos distinguir entre conducta intencional y conducta social; entre conducta social colectiva y conducta individual dentro de un colectivo social; entre hacer algo en atención a algo más y hacer algo en atención a sí mismo. Quizá lo más difícil de todo, necesitamos dar cuenta de las secuencias melódicas de la conducta a través del paso del tiempo. Las actividades humanas, después de todo, no se parecen a una serie de instantáneas fijas, sino que se parecen algo más a la película de nuestra vida.

No puedo esperar responder a todas estas cuestiones, pero espero que al final lo que haya dicho se parezca a una explicación de sentido común de la estructura de la acción. Si estoy en lo cierto, lo que yo digo parecería obviamente correcto. Pero históricamente lo que yo pienso que es la explicación de sentido común no ha parecido obvio. En primer lugar, la tradición conductista en filosofía y en psicología ha llevado a mucha gente a olvidar el componente mental de las acciones. Los con-

ductistas querían definir las acciones, y por cierto toda nuestra vida mental, en términos de puros movimientos físicos. Alguien caracterizó una vez el enfoque conductista, con toda justificación desde mi punto de vista, como anestesia fingida. El extremo opuesto en filosofía ha sido decir que los únicos actos que realizamos alguna vez son los actos mentales internos de volición. Según este punto de vista, estrictamente hablando, nunca levantamos nuestros brazos. Todo lo que hacemos es 'tener la volición' de que nuestros brazos se levanten. Si se levantan, buena suerte, pero no se trata realmente de nuestra acción.

Otro problema es que hasta tiempos recientes la filosofía de la acción era un tema hasta cierto punto olvidado. La tradición occidental ha subrayado persistentemente el conocer como algo más importante que el hacer. La teoría del conocimiento y del significado ha sido más central para sus intereses que la teoría de la acción. Quiero ahora intentar exponer juntamente los aspectos tanto mentales como físicos.

La mejor manera de dar una explicación de la estructura de la conducta es enunciando una serie de principios. Esos principios explicarían los aspectos tanto mentales como físicos de la acción. Al presentarlos no querría discutir de dónde vienen nuestras creencias, deseos y todo lo demás. Pero explicaré cómo figuran en nuestra conducta.

Pienso que la manera más fácil de transmitir esos principios es enunciarlos y, a continuación, intentar defenderlos. Helos aquí:

PRINCIPIO 1. *Las acciones constan característicamente de dos componentes, un componente mental y un componente físico.*

Piénsese, por ejemplo, en empujar un coche. Por un lado, hay ciertas experiencias conscientes de esfuerzo cuando se empuja. Si se tiene éxito, esas experiencias resultarán en el movimiento del cuerpo y en el correspondiente movimiento del coche. Si no se tiene éxito, se habrá tenido al menos el componente mental, esto es, se

habrá tenido, con todo, una experiencia de intentar mover el coche con al menos alguno de los componentes físicos. Habrá tenido tensión muscular, la sensación de presión contra el coche, y así sucesivamente. Esto lleva al

PRINCIPIO 2. *El componente mental es una intención.*

Tiene intencionalidad —es sobre algo. Determina lo que cuenta como un éxito o como un fallo en la acción; y si tiene éxito causa el movimiento corporal que a su vez causa los demás movimientos, tales como el movimiento del coche, que constituye el resto de la acción. En términos de la teoría de la intencionalidad que acabamos de bosquejar, la acción consta de dos componentes, un componente mental y un componente físico. Si tiene éxito el componente mental causa el componente físico. A esta forma de causación la llama 'causación intencional'.

La mejor manera de ver la naturaleza de los diferentes componentes de una acción es tomar por separado cada uno de ellos y examinarlos individualmente. Y de hecho, en un laboratorio, es bastante fácil hacer esto. En neurofisiología tenemos ya experimentos, hechos por Wilder Penfield, de Montreal, donde estimulando eléctricamente una cierta porción del córtex motor del paciente, Penfield podía causar el movimiento de los miembros de éste. Ahora bien, los pacientes se mostraban invariablemente sorprendidos por esto, y característicamente decían cosas tales como: 'Yo no hice eso, lo hizo usted'. En tal caso, hemos tomado por separado el movimiento corporal sin la intención. Obsérvese que en tales casos los movimientos corporales podrían ser los mismos que son en una acción intencional, pero parece completamente claro que hay una diferencia. ¿Cuál es la diferencia? Bien, tenemos también experimentos que se remontan a William James, donde podemos separar el componente mental sin el correspondiente componente físico de la acción. En el caso de James se anestesia el brazo de un paciente, y se le mantiene pegado al costado en una habitación oscura, y a continuación se le ordena levantarlo. El hace lo que él piensa que es obe-

decer la orden, pero más tarde se muestra totalmente sorprendido al descubrir que su brazo no se ha levantado. Ahora bien, en este caso tomamos por separado el componente mental, es decir: la intención, del movimiento corporal. Esto es, podemos decir verdaderamente de él que genuinamente intentó mover su brazo.

Normalmente esos dos componentes van juntos. Usualmente tenemos la intención y el movimiento corporal, pero no son independientes. Lo que intentan articular nuestros dos principios es cómo se relacionan. El componente mental tiene, como parte de sus condiciones de satisfacción, tanto que *representar* como que *causar* el componente físico. Obsérvese de pasada que tenemos un vocabulario bastante extenso de 'intentar' y 'tener éxito', y 'fracasar', de 'intencional', de 'acción' y 'movimiento', para describir el funcionamiento de esos principios.

PRINCIPIO 3. *El género de causación que es esencial tanto a la estructura de la acción como a la explicación de la acción es la causación intencional.*

Los movimientos corporales de nuestras acciones están causados por nuestras intenciones. Las intenciones son causales porque hacen que las cosas sucedan; pero también tienen contenidos y así pueden figurar en el proceso de razonamiento lógico. Pueden ser causales y tener a la vez rasgos lógicos porque el género de causación sobre el que estamos hablando es causación mental o causación intencional. Y en la causación intencional los contenidos mentales afectan al mundo. Todo el aparato funciona porque está realizado en el cerebro, en el modo que expliqué en el primer capítulo.

La forma de causación que estamos discutiendo aquí es completamente diferente de la forma estándar de causación tal como se describe en los libros de texto filosóficos. No es un asunto de regularidades o de leyes encubiertas o de conjunciones constantes. De hecho, pienso que está mucho más cerca de nuestra noción de causación de sentido común, donde solamente queremos decir que algo hace que algo más suceda. Lo que es especial

en la causación intencional es que se trata de un caso de un estado mental que hace que suceda algo más, y ese algo más es el mismo estado de cosas representado por el estado mental que lo causa.

PRINCIPIO 4. *En la teoría de la acción existe una distinción fundamental entre esas acciones que son premeditadas, que son un resultado de algún género de planificación anticipada, y aquellas acciones que son espontáneas, donde hacemos algo sin ninguna reflexión anterior.*

Y correspondiendo a esta distinción necesitamos una distinción entre *intenciones anteriores,* esto es, intenciones formadas antes de la realización de una acción, e *intenciones en la acción,* que son las intenciones que tenemos mientras estamos realizando efectivamente una acción.

Un error común en la teoría de la acción es suponer que todas las acciones intencionales son el resultado de algún tipo de deliberación, que son el producto de una cadena de razonamiento práctico. Pero obviamente, muchas cosas de las que hacemos no se parecen a esto. Nosotros simplemente hacemos algo sin ninguna reflexión anterior. Por ejemplo, en una conversación normal, no se reflexiona sobre lo que se va a decir a continuación, simplemente se dice. En tales casos hay ciertamente una intención, pero no es una intención formada anteriormente a la realización de la acción. Es lo que llamamos una intención en la acción. En otros casos, sin embargo, nos formamos intenciones anteriores. Reflexionamos sobre lo que queremos y lo que es la mejor manera de lograrlo. Este proceso de reflexión (Aristóteles lo llamó 'razonamiento práctico') tiene como resultado de modo característico o bien la formación de una intención anterior o, como Aristóteles señaló también, algo que tiene como resultado la acción misma.

PRINCIPIO 5. *La formación de intenciones anteriores es, al menos de manera general, el resultado práctico. El razonamiento práctico es siempre razonamiento sobre cómo decidir mejor entre deseos en conflicto.*

La fuerza motriz que está detrás de la mayor parte de la acción humana (y animal) es el deseo. Las creencias funcionan solamente para capacitarnos para calcular cómo satisfacer mejor nuestros deseos. Así, por ejemplo, yo quiero ir a París, y creo que la mejor manera, una vez consideradas todas las cosas, es ir en avión, de modo que me formo la intención de ir en avión. Esto es un ejemplo típico de sentido común de razonamiento práctico. Pero el razonamiento práctico difiere esencialmente del razonamiento teórico, del razonamiento sobre lo que es el caso, en que el razonamiento práctico siempre es sobre cómo decidir mejor entre diversos deseos conflictivos que tenemos. Así, por ejemplo, supongamos que quiero ir a París, y resuelvo que la mejor manera es ir en avión. Sin embargo, no hay manera alguna de que pueda hacer esto sin frustrar un buen número de otros deseos que tengo. No quiero gastar dinero, no quiero hacer cola en los aeropuertos, no quiero sentarme en los asientos de los aviones, no quiero que la gente ponga sus codos cuando estoy intentando poner el mío, y así indefinidamente. Sin embargo, a pesar de todos los deseos que frustraré si voy a París en avión, puedo con todo razonar que lo mejor, una vez consideradas todas las cosas, es ir a París en avión. Esto no es solamente típico del razonamiento práctico, sino que pienso que es algo universal en el razonamiento práctico que el razonamiento práctico se ocupe de la adjudicación de deseos en conflicto.

El cuadro que emerge a partir de estos cinco principios es, entonces, que la energía mental que alimenta la acción es una energía que funciona por medio de la causación intencional. Se trata de una forma de energía mediante la cual la causa, ya sea en forma de deseos o de intenciones, representa el mismo estado de cosas que causa.

Volvamos ahora a alguno de aquellos puntos sobre la acción que observamos al principio, porque pienso que hemos reunido piezas suficientes para explicarlos. Hemos observado que las acciones tenían descripciones preferenciales y que, de hecho, el sentido común nos capacitaba para identificar cuáles eran las descripciones prefe-

renciales de las acciones. Ahora podemos ver que la descripción preferencial de una acción está determinada por la intención. Lo que la persona está realmente haciendo, o al menos lo que está intentando hacer, es enteramente un asunto de cuál es la intención con la que está actuando. Por ejemplo, yo sé que estoy intentando llegar al Retiro y no estoy intentando acercarme a la Patagonia, puesto que esa es la intención con la que estoy dando el paseo. Y sé esto sin *observación* alguna, puesto que el conocimiento en cuestión no es conocimiento de mi conducta externa, sino de mis estados mentales internos.

Esto explica además alguno de los rasgos lógicos de las explicaciones que damos de la acción humana. Explicar una acción es dar sus causas. Sus causas son estados psicológicos. Esos estados se relacionan con la acción o bien constituyendo pasos en el razonamiento práctico que llevan a las intenciones o siendo las intenciones mismas. El rasgo más importante de la explicación de la acción merece la pena, sin embargo, enunciarlo como un principio separado; así, pues, llamémoslo

PRINCIPIO 6. *La explicación de una acción tiene que tener el mismo contenido que estaba en la cabeza de la persona cuando realizaba la acción o cuando razonaba hacia su intención de realizar la acción. Si la explicación es realmente explicativa, el contenido que causa la conducta por medio de la causación intencional tiene que ser idéntico al contenido de la explicación de la conducta.*

En este aspecto las acciones difieren de otros eventos del mundo y, correspondientemente, sus explicaciones son diferentes. Cuando explicamos un terremoto o un huracán, el contenido de la explicación sólo tiene que representar lo que sucedió y por qué sucedió. No tiene que causar el evento mismo. Pero al explicar la acción humana, tanto la causa como la explicación tienen contenidos y la explicación sólo explica porque tiene el mismo contenido que la causa.

Hasta aquí hemos estado hablando como si la gente

tuviera sólo intenciones caídas del cielo. Pero, desde luego, esto es muy irreal. Y ahora necesitamos introducir algunas complejidades que acercarán un poco nuestro análisis a los asuntos de la vida real. Nadie tiene jamás sólo una intención, una cualquiera, por sí misma. Por ejemplo, yo tengo intención de ir conduciendo de Madrid a Segovia: puedo tener esta intención de manera completamente espontánea, pero, sin embargo, tengo que tener, con todo, una serie de otros estados intencionales. Tengo que tener una creencia de que tengo un coche y una creencia de que Segovia está a una distancia que se puede cubrir conduciendo. Además, tendré característicamente un deseo de que no haya demasiada circulación en las carreteras y de que el tiempo no sea demasiado malo para conducir. También (y aquí nos acercamos un poco más a la noción de explicación de la acción) característicamente no iré solamente conduciendo hasta Segovia, sino que ire conduciendo hasta Segovia con algún propósito. Y si esto es así, llevaré a cabo razonamientos prácticos —esa forma de razonamiento que no lleva a creencias o conclusiones de argumentos, sino a intenciones y a conducta efectiva. Y cuando entendamos esta forma de razonamiento, habremos dado un gran paso hacia la compresión de la explicación de las acciones. Llamemos a los otros estados intencionales que dan a mi estado intencional el significado particular que tiene, llamémoslos a todos 'la malla de la intencionalidad'. Y podemos decir por medio de una conclusión general —llamemos a esto

PRINCIPIO 7. *Cualquier estado intencional solamente funciona como parte de una malla de otros estados intencionales. Y por 'funciona' quiero decir aquí que sólo determina sus condiciones de satisfacción de manera relativa a toda una porción de otros estados intencionales.*

Ahora bien, cuando comenzamos a indagar los detalles de la malla, descubrimos otro fenómeno interesante. Y es simplemente que las actividades de nuestra mente

no pueden consistir en estados mentales, por así decirlo, pura y simplemente. Más bien, nuestros estados mentales funcionan solamente del modo en que lo hacen en contraste con un *background* de capacidades, habilidades, destrezas, hábitos, maneras de hacer cosas, y posiciones generales hacia el mundo que no consisten ellas mismas en estados intencionales. Para que pueda formarme la intención de ir conduciendo hasta Segovia tengo que tener la capacidad de conducir. Pero la capacidad de conducir no consiste en toda una porción de otros estados intencionales. El ser capaz de conducir abarca más que un puñado de creencias y deseos. Tengo que tener efectivamente la destreza de hacerlo. Este es un caso donde mi saber cómo no es solamente un asunto de saber qué. Llamemos al conjunto de destrezas, hábitos habilidades, etc., en contraste con el cual funcionan los estados intencionales, 'el *background* de la intencionalidad', y a la tesis de la malla; a saber: que cualquier estado intencional solamente funciona como parte de una malla, añadiremos la tesis del *background* —llamémosla

PRINCIPIO 8. *La malla total de la intencionalidad solamente funciona en contraste con un background de capacidades humanas que no son, ellas mismas, estados mentales.*

He dicho que muchas explicaciones supuestamente científicas de la conducta intentan escapar de, o sobrepasar este modelo de sentido común que he estado bosquejando. Pero al final no hay manera alguna, pienso, de que puedan hacer esto, puesto que esos principios no describen solamente los fenómenos: ellos mismos forman parcialmente los fenómenos. Considérense, por ejemplo, las explicaciones freudianas. Cuando Freud está haciendo metapsicología, esto es, cuando está dando la teoría de lo que está haciendo, usa a menudo comparaciones científicas. Hay una gran cantidad de analogías entre la psicología, y el electromagnetismo o la hidráulica, y hemos de pensar en la mente como si funcionase de acuerdo con la analogía de los principios

hidráulicos, y así sucesivamente. Pero cuando él está examinando efectivamente a un paciente, y está describiendo efectivamente la naturaleza de la neurosis de algún paciente, es sorprendente hasta qué punto las explicaciones que de él da son explicaciones de sentido común. Dora se comporta de la manera que lo hace porque está enamorada de Herr K, o porque está imitando a su primo que se ha marchado a Mariazell. Lo que Freud añade al sentido común es la observación de que a menudo los estados mentales que causan nuestra conducta son inconscientes. De hecho, están reprimidos. A menudo ofrecemos resistecia a admitir que tenemos ciertos estados mentales porque nos avergonzamos de ellos, o por alguna otra razón. Y en segundo lugar, él añade una teoría de la transformación de los estados mentales, cómo un tipo de estado intencional puede transformarse en otro. Pero con la adición de ésta y otras excrecencias, la forma freudiana de explicación es la misma que las formas de sentido común. Sugiero que el sentido común es muy probable que persista incluso a medida que adquiramos otras explicaciones más científicas de la conducta. Puesto que la estructura de la explicación tiene que acoplarse a la estructura de los fenómenos explicados, las mejoras en la explicación no es probable que tengan estructuras nuevas e inauditas.

En este capítulo he intentado explicar cómo y en qué sentido la conducta contiene y a la vez es causada por estados mentales internos. Puede parecer sorprendente que gran parte de la psicología y de la ciencia cognitiva hayan intentado negar esas relaciones. En el próximo capítulo voy a explorar alguna de las consecuencias de mi punto de vista sobre la conducta humana para las ciencias sociales. ¿Por qué las ciencias sociales han sufrido los fallos que han sufrido y alcanzado el éxito que han alcanzado, y qué podemos esperar aprender de ellas razonablemente?

5

PERSPECTIVAS PARA LAS CIENCIAS SOCIALES

En este capítulo quiero discutir uno de los problemas intelectuales más debatidos de la época presente: ¿Por qué los métodos de las ciencias naturales no nos han dado el género de rendimiento en el estudio de la conducta humana que han dado en física y en química? ¿Y en qué tipo de ciencias 'sociales' o 'de la conducta' podemos en cualquier caso esperar razonablemente? Voy a sugerir que hay ciertas diferencias radicales entre la conducta humana y los fenómenos estudiados en las ciencias naturales. Argumentaré que esas diferencias dan cuenta tanto de los fallos como de los éxitos que hemos tenido en las ciencias humanas.

Al principio quiero llamar la atención sobre una diferencia importante entre la forma de las explicaciones de la conducta humana de sentido común y la forma estándar de la explicación científica. De acuerdo con la teoría estándar de la explicación científica, explicar un fenómeno consiste en mostrar cómo su ocurrencia se sigue de ciertas leyes científicas. Esas leyes son generalizaciones universales sobre cómo suceden las cosas. Por ejemplo, si a uno se le da un enunciado de las leyes relevantes que describen la conducta de un cuerpo que está cayendo, y uno sabe de dónde ha partido, uno puede deducir efectivamente lo que le sucederá al cuerpo en cuestión. Igualmente, si se quiere explicar una ley, se puede deducir de alguna ley de nivel superior.

De acuerdo con esto la explicación y la predicción son perfectamente simétricas. Se predice deduciendo lo que sucederá; se explica deduciendo lo que ha sucedido. Ahora bien, sea cual sea el mérito que este tipo de explicación pueda tener en las ciencias naturales, una de las cosas que quiero subrayar en este capítulo es que carece completamente de valor para explicar la conducta humana. Y esto no sucede porque carezcamos de leyes para explicar ejemplos individuales de conducta humana. Es así porque incluso si tuviésemos tales leyes, nos serían a pesar de todo inútiles. Pienso que se puede lograr fácilmente que ustedes vean esto pidiéndoles que imaginen cuál sería la situación si tuviésemos efectivamente una 'ley', esto es, una generalización universal, concerniente a algún aspecto de su conducta.

Supongamos que en las últimas elecciones usted votó por los conservadores, y supongamos que usted votó por los conservadores porque pensaba que harían más cosas para resolver el problema de la inflación que ninguno de los demás partidos. Ahora bien, supongamos que esto es solamente un hecho puro y simple sobre por qué votó por los conservadores, como es igualmente un hecho puro y simple que usted votó efectivamente por los conservadores. Supongamos además que algunos sociólogos políticos proponen una generalización universal, absolutamente carente de excepciones, sobre la gente que encaja exactamente en su descripción —su estatus socio-económico, su nivel de ingresos, su educación, sus otros intereses, y así sucesivamente. Supongamos que la generalización absolutamente carente de excepciones es al efecto de que la gente semejante a usted vota invariablemente por los conservadores. Ahora bien, yo quiero preguntar. ¿Qué explica el por qué usted votó por los conservadores? ¿Es la razón que usted acepta sinceramente? ¿O la generalización universal? Quiero argumentar que nosotros jamás aceptaríamos la generalización como explicación de nuestra propia conducta. La generalización enuncia una regularidad. El conocimiento de tal regularidad puede ser útil para la predicción, pero no explica nada sobre los casos individuales de conducta humana. De hecho invita a una explicación posterior.

Por ejemplo, ¿por qué toda la gente de ese grupo votó por los conservadores? Una respuesta se sugiere a sí misma. Usted votó por los conservadores porque está preocupado por la inflación —quizás la gente de su grupo estaba particularmente afectada por la inflación y este es el porqué todos ellos votaron de la misma manera.

Dicho brevemente, no aceptamos una generalización como explicación de nuestra conducta ni de la de ningún otro. Y si se hallase una generalización, ella misma requeriría una explicación en la que nosotros estuviéramos detrás en primer lugar. Y por lo que respecta a la conducta humana, el tipo de explicación que normalmente buscamos es aquél que especifica los estados mentales— creencias, temores, esperanzas, deseos y así sucesivamente— que funcionan causalmente en la producción de la conducta, del modo que he descrito en el capítulo anterior.

Volvamos a nuestra pregunta original: ¿Por qué no parece que tengamos leyes de las ciencias sociales en el sentido en que tenemos leyes de las ciencias naturales? Hay varias respuestas estándar a esta cuestión. Algunos filósofos señalan que no tenemos una ciencia de la conducta por la misma razón que no tenemos una ciencia de los muebles. No podríamos tener una tal ciencia porque no hay ningún rasgo físico que tengan en común las sillas, las mesas, los pupitres y otros muebles que los capacite para caer bajo un conjunto común de leyes de los muebles. Y además no necesitamos realmente tal ciencia porque cualquier cosa que necesitamos explicar —por ejemplo, ¿por qué son sólidas las mesas de madera? o ¿por qué se oxidan los muebles de hierro?— puede ser explicado por las ciencias ya existentes. De igual modo no hay rasgo alguno que todas las conductas humanas tengan en común. Y, además, las cosas particulares que deseamos explicar pueden ser explicadas por la física, y la fisiología, y todo el resto de las ciencias existentes.

En un argumento relacionado con lo anterior algunos filósofos señalan que quizás nuestros conceptos para describirnos a nosotros mismos y a otros seres humanos no se acoplan de manera correcta a los conceptos de cien-

cias básicas tales como la física y la química. Quizás, ellos sugieren, la ciencia humana es parecida a la ciencia del tiempo. Tenemos una ciencia del tiempo, la meteorología, pero no es una ciencia estricta porque las cosas que nos interesan sobre el tiempo no se acoplan a las categorías *naturales* que tenemos para la física. Conceptos del tiempo meteorológico tales como 'claros en el centro' o 'parcialmente nuboso en Asturias' no están relacionados sistemáticamente con los conceptos de la física. Una poderosa expresión de este punto de vista está en la obra de Jerry Fodor. Él sugiere que las ciencias especiales como la geología o la meteorología son sobre rasgos del mundo que pueden realizarse en la física en una diversidad de maneras y que esta conexión laxa entre la ciencia especial y la ciencia más básica de la física es también característica de las ciencias sociales. Lo mismo que las montañas y las tormentas pueden realizarse en diferentes géneros de estructuras microfísicas, así también el dinero puede realizarse como oro, plata o papel impreso. Y tales conexiones disyuntivas entre los fenómenos de más alto nivel y los de nivel más bajo nos permiten efectivamente tener ciencias ricas, pero no permiten leyes *estrictas,* puesto que la forma de las conexiones laxas permitirá leyes que tienen excepciones.

Otro argumento a favor del punto de vista de que no podemos tener leyes estrictas que conecten lo mental y lo físico está en la afirmación de Donald Davidson de que los conceptos de racionalidad, consistencia y coherencia son parcialmente constitutivos de nuestra noción de fenómeno mental; y esas nociones no se relacionan sistemáticamente con las nociones de la física. Como Davidson dice, no tienen 'eco' en física. Una dificultad que tiene este punto de vista es que hay montones de ciencias que contienen nociones constitutivas que, similarmente, no tienen ningún eco en física, pero que a pesar de todo son ciencias bastante sólidas. La biología, por ejemplo, exige el concepto de organismo, y 'organismo' no tiene ningún eco en física, pero la biología no deja por ello de ser una ciencia dura.

Otro punto de vista, ampliamente mantenido, es que las interrelaciones complejas de nuestros estados men-

tales nos impiden lograr siempre un conjunto sistemático de leyes que los conecten a los estados neurofisiológicos. De acuerdo con este punto de vista, los estados mentales van en redes complejas, interrelacionadas, y de este modo no pueden ponerse sistemáticamente en correspondencia con tipos de estados cerebrales. Pero una vez más, este argumento es inconcluyente. Supongamos que, por ejemplo, Noam Chomsky tiene razón al pensar que cada uno de nosotros tiene un conjunto complejo de reglas de gramática universal programadas en su cerebro en el momento del nacimiento. No hay nada en la complejidad o interdependencia de las reglas que impida que estén sistemáticamente realizadas en la neurofisiología del cerebro. La interdependencia y la complejidad por sí mismas no son un argumento suficiente contra la posibilidad de leyes psicofísicas estrictas.

Encuentro todas esas explicaciones sugerentes pero no creo que capten adecuadamente las diferencias realmente radicales entre las ciencias físicas y las mentales. La relación entre la sociología y la economía de un lado y la física de otro es en realidad distinta por completo de las relaciones de, por ejemplo, la meteorología, la geología y la biología y otras ciencias naturales especiales con la física; y necesitamos intentar enunciar cómo son exactamente. Idealmente, me gustaría ser capaz de dar, paso a paso, un argumento que mostrase las limitaciones que existen sobre las posibilidades de ciencias sociales estrictas y mostrar con todo la naturaleza real y el poder de esas disciplinas. Pienso que debemos abandonar de una vez por todas la idea de que las ciencias sociales están en un estado semejante a la física antes de Newton, y que aquéllo por lo que estamos esperando es un conjunto de leyes newtonianas de la mente y de la sociedad.

En primer lugar ¿en qué se supone que consiste exactamente el problema? Podría decirse, 'Seguramente los fenómenos sociales y psicológicos son tan reales como cualquier otra cosa. Así, pues, ¿por qué no puede haber leyes de su conducta?' ¿Por qué habría de haber leyes de la conducta de las moléculas pero no leyes de la conducta de las sociedades? Bien, una manera de deshacerse de una tesis es imaginar que es verdadera y mostrar a con-

tinuación que esa suposición es de alguna manera absurda. Supóngase que tuviésemos efectivamente leyes de la sociedad y leyes de la historia que nos capacitasen para predecir cuándo habría guerras y revoluciones. Supóngase que pudiésemos predecir guerras y revoluciones con la misma precisión y exactitud con que podemos predecir la aceleración de un cuerpo que cae en el vacío al nivel del mar.

El problema real es este: cualquier cosa que sean las guerras y revoluciones, incluyen montones de movimientos moleculares. Pero eso tiene la consecuencia de que cualquier ley estricta sobre guerras y revoluciones tendría que acoplarse perfectamente con las leyes sobre movimientos moleculares. Para que una revolución estallase tal y tal día, las moléculas relevantes tendrían que estar moviéndose en la dirección correcta. Pero si esto es así, entonces les leyes que predicen la revolución tendrán que hacer las mismas predicciones al nivel de las revoluciones y sus participantes que las leyes de los movimientos moleculares hacen al nivel de las partículas físicas. Así, nuestra pregunta original puede reformularse, ¿por qué no pueden las leyes al más alto nivel, el nivel de las revoluciones, acoplarse perfectamente con las leyes que están al nivel más bajo, el nivel de las partículas? Bien, para ver por qué no pueden, examinemos algunos casos donde hay realmente un acoplamiento perfecto entre las leyes de orden más elevado y las leyes de nivel más bajo, y a continuación podemos ver cómo difieren esos casos de los casos sociales.

Uno de los éxitos mayores de todos los tiempos en la reducción de las leyes de un nivel a las de un nivel inferior es la reducción de las leyes de los gases —la ley de Boyle y la ley de Charles— a las leyes de la mecánica estadística. ¿Cómo funciona la reducción? Las leyes de los gases se ocupan de la relación entre la presión, temperatura y volumen de los gases. Predicen, por ejemplo, que si se incrementa la temperatura de un gas que está en un cilindro, se incrementará la presión sobre las paredes del cilindro. Las leyes de la mecánica estadística se ocupan de la conducta de las masas de pequeñas partículas. Predicen, por ejemplo, que si se incrementa la

proporción del movimiento de las partículas de un gas, cada vez más partículas golpearán las paredes del cilindro y cada vez las golpearán más fuertemente. La razón por la que se obtiene un acople perfecto entre esos dos conjuntos de leyes es que la explicación de la temperatura, presión y volumen puede darse enteramente en términos de la conducta de las partículas. Al incrementar la temperatura del gas se incrementa la velocidad de las partículas, y al incrementar el número y la velocidad de las partículas que golpean el cilindro se incrementa la presión. Se sigue que un incremento en la temperatura producirá un incremento en la presión. Supongamos ahora por mor del argumento que la situación no era semejante a esta. Supongamos que no hubiera ninguna explicación de la presión y la temperatura en términos de la conducta de las partículas más fundamentales. Entonces, cualesquiera leyes al nivel de la presión y la temperatura serían milagrosas, puesto que sería milagroso que el modo en que la presión y la temperatura estuviesen actuando coincidiese exactamente con el modo en que las partículas estaban actuando, si no había relación sistemática alguna entre la conducta del sistema al nivel de la presión y la temperatura, y la conducta del sistema al nivel de las partículas.

Este ejemplo es un caso muy simple. Así, pues, tomemos un ejemplo ligeramente más complejo. Se trata de una ley de la 'ciencia de la nutrición' que dice que la ingestión calórica es igual al rendimiento calórico, más o menos el depósito de grasa. No es quizá una ley muy imaginativa, pero, a pesar de todo, es bastante realista. Tiene la consecuencia conocida por la mayor parte de nosotros de que si se come mucho y no se hace suficiente ejercicio, se engorda. Ahora bien, esta ley, a diferencia de lo que sucede con las leyes de los gases, no está fundada de ninguna manera simple en la conducta de las partículas. El fundamento no es simple —puesto que, por ejemplo, hay una serie más bien compleja de procesos por medio de los cuales la comida se convierte en depósitos de grasa en los organismos vivientes. Sin embargo, hay con todo un fundamento —aunque complejo— de esta ley en términos de la conducta de partícu-

las más fundamentales. Permaneciendo las demás cosas igual, cuando se come mucho, las moléculas estarán moviéndose exactamente en la dirección correcta para hacer que se engorde.

Podemos ahora argumentar adicionalmente hacia la conclusión de que no habrá ley alguna de las guerras y las revoluciones de un modo en el que hay leyes de los gases y de la nutrición. Los fenómenos del mundo que seleccionamos con conceptos tales como guerra y revolución, matrimonio, dinero y propiedad no están fundados sistemáticamente en la conducta de elementos que están al nivel más básico de un modo en que los fenómenos que seleccionamos con conceptos como depósito de grasa y presión están fundados sistemáticamente en la conducta de elementos que están al nivel más básico. Obsérvese que es este tipo de fundamentación la que nos capacita característicamente para hacer avances mayores en los niveles más elevados de una ciencia. La razón de que el descubrimiento de la estructura del ADN sea tan importante para la biología, o que la teoría microbiana de la enfermedad sea tan importante para la medicina, es que en cada caso se mantiene firmemente la promesa de explicar sistemáticamente características de niveles más elevados, tales como los rasgos hereditarios y los síntomas de enfermedades, en términos de elementos más fundamentales.

Pero ahora surge esta pregunta: ¿Si los fenómenos sociales y psicológicos no están fundamentados de esta manera, por qué no lo están? ¿Por qué no podrían estarlo? Dando por sentado que no están así fundamentados, ¿por qué no? Esto es, guerras y revoluciones, al igual que todo lo demás, consisten en movimientos de moléculas. Así, pues, ¿por qué no pueden los fenómenos sociales tales como guerras y revoluciones estar relacionados sistemáticamente con los movimientos de moléculas de la misma manera que las relaciones entre ingestiones calóricas y depósitos de grasa son sistemáticas?

Para ver por qué esto no puede ser así hemos de preguntar qué rasgos tienen los fenómenos sociales que nos capacitan para aglutinarlos en categorías. ¿Cuáles son los principios fundamentales de acuerdo con los cuales cate-

gorizamos los fenómenos psicológicos y sociales. Un rasgo esencial es éste: para buen número de fenómenos sociales y psicológicos el concepto que nombra el fenómeno es, él mismo, un constituyente del fenómeno. Para que algo cuente como una ceremonia de matrimonio o un sindicato, o una propiedad, o dinero, o incluso una guerra o revolución, la gente incluída en esas actividades tiene que tener ciertos pensamientos apropiados. Así, por ejemplo, tienen que pensar que eso es lo que es. Así, por ejemplo, para casarse o comprar una propiedad usted y otras personas tienen que pensar que es eso lo que están haciendo. Ahora bien, esta característica es crucial para los fenómenos sociales. Pero no hay nada parecido en las ciencias biológicas y físicas. Algo puede ser un árbol o una planta, o alguna persona puede tener tuberculosis incluso si nadie piensa. 'He aquí un árbol, o una planta, o un caso de tuberculosis', e incluso si nadie piensa sobre ello en absoluto. Pero muchos de los términos que describen los fenómenos sociales tienen que entrar en su constitución. Y esto tiene el resultado 'de que tales términos tienen un género peculiar de auto-referencialidad.' 'Dinero' se refiere a cualquier cosa que la gente usa como, y piensa que es, dinero. 'Promesa' se refiere a cualquier cosa que la gente intenta que sea, y considera que es, una promesa. No estoy diciendo que para tener la institución del dinero la gente tenga que tener esa misma palabra, o algún sinónimo exacto, en su vocabulario. Más bien, tienen que tener ciertos pensamientos y actitudes sobre algo para que eso cuente como dinero y esos pensamientos y actitudes son parte de la misma definición de dinero.

Hay otra consecuencia crucial de este rasgo. El principio definidor de tales fenómenos sociales no establece ningunos límites físicos de ningún tipo sobre lo que puede contar como su realización física. Y esto significa que no puede haber ninguna conexiones sistemáticas entre las propiedades físicas y las propiedades sociales o mentales del fenómeno. Los rasgos sociales en cuestión están determinados en parte por las actitudes que tomamos hacia ellos. Las actitudes que tomamos hacia ellos no están constreñidas por los rasgos físicos de los fenómenos en

cuestión. Además, no puede haber un acople del nivel mental y del nivel de la física del tipo que sería necesario para hacer posibles las leyes estrictas de las ciencias sociales.

El paso principal del argumento a favor de una discontinuidad entre las ciencias sociales y las ciencias naturales depende del carácter mental de los fenómenos sociales. Y es este rasgo el que todas esas analogías que he mencionado antes —esto es, entre la meteorología, la biología y la geología— olvidan. La discontinuidad radical entre las disciplinas sociales de un lado y las ciencias naturales de otro deriva del papel de la mente en esas disciplinas.

Considérese la afirmación de Fodor de que las leyes sociales tendrán que tener excepciones, puesto que los fenómenos al nivel social están en correspondencia de manera vaga o disyuntiva con los fenómenos físicos. Una vez más esto no da cuenta de las discontinuidades radicales sobre las que he estado llamando la atención. Incluso si este tipo de disyunción fuese verdadera hasta cierto punto, le está siempre abierta a la próxima persona la posibilidad de añadirle indefinidamente muchas maneras. Supóngase que el dinero ha tomado siempre un rango limitado de formas físicas: oro, plata, y papel impreso, por ejemplo. Con todo, aún le está abierta a la próxima persona o sociedad la posibilidad de tratar alguna cosa distinta como dinero. Y de hecho la realización física no importa para las propiedades del dinero con tal de que la realización física permita que esa materia se use como un medio de cambio.

'Bien', alguien podría objetar, 'para tener ciencias sociales rigurosas no necesitamos un acople estricto entre propiedades de las cosas del mundo. Todo lo que necesitamos es un acople estricto entre propiedades psicológicas y rasgos del cerebro. La fundamentación real de la economía y de la sociología en el mundo físico no está en las propiedades de los objetos que encontramos en torno nuestro, está en las propiedades físicas del cerebro. Así, incluso si pensar que algo es dinero es esencial para el hecho de que sea dinero, con todo pensar que algo es dinero puede ser, y efectivamente según nuestra

explicación, es un proceso del cerebro. Así, para mostrar que no puede haber leyes estrictas de las ciencias sociales se tiene que mostrar que no puede haber correlación estricta alguna entre tipos de estados mentales y tipos de estados cerebrales, y esto no se ha mostrado'.

Para ver por qué no puede haber tales leyes, examinemos algunas áreas donde parece probable que vayamos a tener una neuropsicología estricta, leyes estrictas que ponen en correlación fenómenos mentales y fenómenos neuropsicológicos. Considérese el dolor. Parece razonable suponer que las causas neuropsicológicas de los dolores, al menos en los seres humanos, son completamente limitadas y específicas. De hecho, discutimos algunas de ellas en un capítulo anterior. No parece haber en principio ningún obstáculo para una perfecta neurofisiología del dolor. Pero ¿qué sucede con, pongamos por caso, la visión? Una vez más es difícil ver en principio ningún obstáculo para obtener una neurofisiología adecuada de la visión. Podríamos incluso alcanzar nuestro objetivo cuando pudiésemos describir perfectamente las condiciones neurofisiológicas para tener ciertos tipos de experiencias visuales. La experiencia de ver que algo, por ejemplo, es rojo. Según mi explicación nada nos impide tener tal psicología.

Pero ahora viene la parte difícil: aunque pudiésemos obtener correlaciones sistemáticas entre neurofisiología y dolor o neurofisiología y la experiencia visual de rojo, no podríamos dar explicaciones similares de la neurofisiología de *ver* que algo era dinero. ¿Por qué no? Dando por sentado que siempre que usted ve que hay algo de dinero delante de usted se produce algún proceso neurofisiológico, ¿qué es lo que impide que sea el mismo proceso siempre? Bien, del hecho de que el dinero puede tener un rango indefinido de formas físicas, se sigue que puede tener un rango indefinido de efectos estimulativos sobre nuestros sistemas nerviosos. Pero puesto que puede tener un rango indefinido de modelos estimulativos sobre nuestros sistemas visuales, constituiría un milagro el que todos ellos produjesen exactamente el mismo efecto neurofisiológico sobre el cerebro.

Y lo que vale para *ver* que algo es dinero vale incluso

más vigorosamente para *creer* que esto es dinero. Sería nada menos que milagroso el que siempre que alguien creyese que andaba mal de dinero, en cualquier lenguaje y cultura que él tuviese esta creencia, tuviese exactamente el mismo tipo de realización neurofisiológica. Y esto sucede simplemente porque el rango de posibles estímulos neurofisiológicos que podría producir esa misma creencia es infinito. Paradójicamente, el modo en que lo mental afecta a lo físico impide el que haya jamás una ciencia estricta de lo mental.

Obsérvese que en los casos en que no tenemos este tipo de interacción entre los fenómenos físicos y sociales, este obstáculo que impide tener ciencias sociales y estrictas no está presente. Considérese el ejemplo que he mencionado anteriormente de la hipótesis de Chomsky de la gramática universal. Supóngase que cada uno de nosotros tiene innatamente programadas en su cerebro las reglas de gramática universal. Puesto que esas reglas están en el cerebro desde el nacimiento e independientemente de cualesquiera relaciones que el organismo tenga con su entorno, no hay nada en mi argumento que impida que haya leyes psicofísicas estrictas que conecten esas reglas con rasgos del cerebro, por muy interrelacionadas que esas reglas puedan estar o por muy complicadas que sean. De nuevo, muchos animales tienen estados mentales conscientes, pero hasta donde sabemos carecen de la autorreferencialidad que va emperajada con el tener lenguajes humanos e instituciones sociales. No hay nada en mi argumento que bloquee la posibilidad de una ciencia de la conducta animal. Por ejemplo, podría haber leyes estrictas que correlacionasen los estados cerebrales de los pájaros con su conducta por lo que respecta a la construcción de nidos.

He prometido darles un bosquejo al menos de un argumento paso a paso. Veamos lo lejos que llego manteniendo mi promesa. Establezcamos el argumento como una serie de pasos.

1. Para que haya leyes de las ciencias sociales en el sentido en que hay leyes de la física tiene que haber alguna correlación sistemática entre fenómenos identificados en términos sociales y psicológicos y fenómenos

identificados en términos físicos. Esto puede ser tan complejo como el modo en que el tiempo meteorológico está conectado con los fenómenos de la física, pero tiene que haber alguna correlación sistemática. Dicho en la jerga contemporánea, tiene que haber algunos principios puente entre las leyes de nivel superior y las leyes de nivel inferior.

2. Los fenómenos sociales están definidos en gran parte en términos de las actitudes psicológicas que la gente toma. Lo que cuenta como dinero, o como una promesa, o como un matrimonio, es en gran parte un asunto de lo que la gente piensa que es dinero o una promesa o un matrimonio.

3. Esto tiene como consecuencia que esas categorías son físicamente abiertas. No hay, estrictamente hablando, ningún límite físico a lo que se puede considerar como, o estipular que es, dinero, o una promesa, o una ceremonia de matrimonio.

4. Esto implica que no puede haber ningunos principios-puente entre los rasgos sociales y físicos del mundo; esto es, entre fenómenos descritos en términos sociales y los mismos fenómenos descritos en términos físicos. No podemos tener ni siquiera el tipo de principios disyuntivos laxos que tenemos para el tiempo meteorológico o la digestión.

5. Además, es imposible obtener el género correcto de principios-puente entre fenómenos descritos en términos mentales y fenómenos descritos en términos neurofisiológicos; esto es, entre el cerebro y la mente. Y esto es así porque hay un rango indefinido de condiciones estimulativas para cualquier concepto social dado. Y este enorme rango impide que los conceptos que no están construidos dentro de nosotros estén realizados de una manera que correlacione sistemáticamente los rangos mentales y físicos.

Quiero concluir este capítulo describiendo lo que me parece el verdadero carácter de las ciencias sociales. Las ciencias sociales en general versan sobre varios aspectos de la intencionalidad. La economía versa sobre la producción y distribución de bienes y servicios. Obsérvese que el economista en su trabajo puede simplemente dar

por sentada la intencionalidad. Supone que los empresarios están intentando hacer dinero y que los consumidores preferirían estar mejor de dinero que estar peor. Y las 'leyes' de la economía enuncian entonces conclusiones o consecuencias de tales suposiciones. Dadas ciertas suposiciones el economista puede deducir que los empresarios racionales venderán cuando sus costes marginales igualen a sus ingresos marginales. Ahora bien, obsérvese que la ley no predice que el hombre de negocios se pregunte a sí mismo: '¿Estoy ya vendiendo cuando los costos marginales igualan a los ingresos marginales?' No, la ley no enuncia el contenido de la intencionalidad individual. La teoría de la empresa en microeconomía extrae las consecuencias de ciertas suposiciones sobre los deseos y posibilidades de los consumidores y la empresas que se ocupan de comprar, producir y vender. La macroeconomía extrae las consecuencias de tales suposiciones para naciones y sociedades enteras. Pero el economista no tiene que preocuparse por cuestiones tales como '¿qué es realmente el dinero?', o '¿qué es realmente un deseo?' Si es muy sofisticado en economía del bienestar puede preocuparse sobre el carácter exacto de los deseos de los empresarios y los consumidores, pero incluso en tal caso la parte sistemática de su disciplina consiste en extraer las consecuencias de hechos sobre la intencionalidad.

Puesto que la economía no está fundamentada en hechos sistemáticos sobre las propiedades físicas tales como la estructura molecular, del modo en que la química está fundada en hechos sistemáticos sobre la estructura molecular, sino que más bien está fundada en hechos sistemáticos sobre la intencionalidad humana, sus deseos, prácticas, estados de tecnología y estados de conocimiento, se sigue que la economía no puede estar libre de la historia o del contexto. La economía como ciencia presupone ciertos hechos históricos sobre la gente y las sociedades que no son ellos mismos parte de la economía. Y cuando esos hechos cambian, la economía tiene que cambiar. Por ejemplo, hasta hace poco la curva de Phillips (una fórmula que pone en relación una serie de factores de las sociedades industrializadas, parecía dar una descripción adecuada de las realidades económicas en esas socieda-

des. Más tarde, no ha funcionado tan bien. La mayor parte de los economistas creen que esto es así porque no describe de manera precisa la realidad. Pero ellos podrían considerar: quizá describía con precisión la realidad tal como era en aquel tiempo. Sin embargo, después de la crisis del petróleo y varios otros eventos de los años setenta la realidad cambió. La economía es una ciencia sistemática formalizada, pero no es independiente del contexto ni está libre de la historia. Está fundada en las prácticas humanas, pero esas prácticas no son ellas mismas atemporales, eternas o inmutables. Si por alguna razón el dinero tuviera que hacerse de hielo, entonces sería una ley estricta de la economía que el dinero se funde a temperaturas superiores a los cero grados centígrados. Pero esta ley funcionaría sólo en la medida en que el dinero estuviese hecho de hielo, y además de esto, no nos dice lo que es interesante para nosotros sobre el dinero.

Volvamos a la lingüística. La aspiración contemporánea estándar de la lingüística es enunciar las diversas reglas —fonológicas, sintácticas y semánticas— que relacionan sonidos y significados en los diversos lenguajes naturales. Una ciencia completamente ideal de la lingüística daría el conjunto completo de reglas para cada lenguaje humano natural. No estoy seguro de que ésta sea la meta correcta de la lingüística ni tan siquiera de que sea una meta que sea posible alcanzar, pero para los presentes propósitos lo importante es observar que una vez más se trata de una ciencia aplicada de la intencionalidad. No es en lo más mínimo algo parecido a la química o a la geología. Se ocupa de especificar aquellos contenidos intencionales históricamente determinados que están en las mentes de los hablantes de varios lenguajes que son efectivamente responsables de la competencia lingüística humana. Y como sucede con la economía, la cola que encuaderna la lingüística es la intencionalidad humana.

El resultado de este capítulo puede enunciarse ahora de manera muy simple. La discontinuidad radical entre las ciencias sociales y las naturales no procede del hecho de que hay solamente una conexión disyuntiva entre los

fenómenos sociales y físicos. No procede ni siquiera del hecho de que las disciplinas sociales tienen conceptos constitutivos que no tienen eco alguno en física ni incluso de la gran complejidad de la vida social. Muchas disciplinas, tales como la geología, la biología y la meteorología tienen esos rasgos, pero esto no les impide ser ciencias naturales sistemáticas. No, la discontinuidad radical deriva del carácter intrínsecamente mental de los fenómenos sociales y psicológicos.

El hecho de que las ciencias sociales estén alimentadas por la mente es la fuente de su debilidad en comparación con las ciencias naturales. Pero es también precisamente la fuente de su fuerza como ciencias sociales. Lo que queremos de las ciencias sociales y lo que obtenemos de las ciencias sociales como mucho son teorías de la intencionalidad pura y aplicada.

6

EL LIBRE ALBEDRÍO

En estas páginas he intentado responder a lo que para mí son las más inquietantes cuestiones sobre cómo nosotros, como seres humanos, encajamos en el universo. Nuestra concepción de nosotros mismos como agentes libres es fundamental para nuestra autoconcepción global. Ahora bien, idealmente me gustaría ser capaz de mantener tanto mis concepciones de sentido común como mis creencias científicas. En el caso de la relación entre la mente y el cuerpo, por ejemplo, fui capaz de hacerlo. Pero cuando se llega a la cuestión de la libertad y del determinismo, no soy capaz —al igual que muchos otros filósofos— de reconciliar las dos.

Habría que pensar que después de más de dos mil años de preocuparse de él, el problema del libre albedrío debería estar ya finalmente resuelto. Bien, efectivamente, muchos filósofos piensan que ha sido resuelto. Piensan que fue resuelto por Thomas Hobbes y David Hume y otros diversos filósofos de inclinación empirista cuyas soluciones han sido repetidas y mejoradas correctamente en el siglo XX. Yo pienso que no se ha resuelto. En esta conferencia quiero proporcionar una explicación de en qué consiste el problema, y de por qué la solución contemporánea no es una solución, y concluir entonces intentando explicar por qué es probable que el problema siga con nosotros.

Por una parte estamos inclinados a decir que puesto

que la naturaleza consta de partículas y sus relaciones entre sí, y puesto que de todo se puede dar cuenta en términos de esas partículas y sus relaciones, no hay lugar simplemente para el libre albedrío. Por lo que respecta a la libertad humana, no importa si la física es determinista, como era la física de Newton, o si se permite una indeterminación al nivel de la física de partículas, como hace la mecánica cuántica contemporánea.

El indeterminismo en física en el nivel de las partículas realmente no sirve de apoyo a ninguna doctrina del libre albedrío, puesto que primero la indeterminación estadística al nivel de las partículas no muestra indeterminación alguna al nivel de los objetos que nos interesan; los cuerpos humanos, por ejemplo. Y en segundo lugar, incluso si hay un elemento de indeterminación en la conducta de las partículas físicas —incluso si sólo son predecibles estadísticamente— con todo esto no da oportunidad alguna para el libre albedrío humano, puesto que del hecho de que las partículas están solamente determinadas estadísticamente no se sigue que la mente humana pueda forzar a las partículas estadísticamente determinadas a desviarse de sus trayectorias. El indeterminismo no constituye prueba alguna de que haya o pueda haber alguna energía mental de la libertad humana que pueda mover las moléculas hacia direcciones hacia las que de otra manera ellas no se hubieran movido. Así, parece realmente como si todo lo que sabemos sobre física nos forzase a una negación de la libertad humana.

La imagen más fuerte para comunicar esta concepción del determinismo es todavía la formulada por Laplace: si un observador ideal conociese las posiciones de todas las partículas en un instante dado y conociese todas las leyes que gobiernan sus movimientos, podría predecir y retrodecir toda la historia del universo. Algunas de las predicciones de la mecánica cuántica contemporánea laplaciana podrían ser estadísticas, pero con todo no harían sitio alguno para el libre albedrío.

Basta con esto por lo que respecta al determinismo. Vayamos ahora al argumento a favor del libre albedrío. Como muchos filósofos han señalado, si hay un hecho de

experiencia con el que todos nosotros estamos familiarizados, es el simple hecho de que nuestras elecciones, decisiones, razonamientos y cogitaciones parecen tener influencia sobre nuestra conducta efectiva. Hay todo tipo de experiencias que tenemos en la vida donde parece pura y simplemente un hecho de nuestra experiencia que aunque hicimos una cosa, tenemos la sensación de que sabemos perfectamente bien que podríamos haber hecho algo distinto. Sabemos que podríamos haber hecho algo distinto porque elegimos una cosa por ciertas razones. Pero éramos conscientes de que había también razones para escoger algo distinto y, de hecho, podríamos haber actuado de acuerdo con esas razones y haber elegido ese algo distinto. Otra manera de expresar este punto consiste en decir: es sólo un puro hecho empírico sobre nuestra conducta el que no es predecible, en el sentido en que la conducta de los objetos que ruedan hacia abajo por un plano inclinado, es predecible. Y la razón por la que no es predecible de ese modo es que a menudo podríamos haber actuado de manera distinta a como de hecho lo hicimos. Si queremos alguna prueba empírica de este hecho, podemos señalar simplemente que siempre somos capaces de falsear cualesquiera predicciones que alguien pudiera tomarse la molestia de hacer sobre nuestra conducta. Si alguien predice que yo voy a hacer algo, yo podría destruir la predicción y hacer algo distinto. Ahora bien, este tipo de opción simplemente no está abierta a los glaciares que se desplazan por las laderas de las montañas. o a las bolas que ruedan por los planos inclinados, o a los planetas que se mueven en sus órbitas elípticas.

Este es un enigma filosófico característico. De un lado, un conjunto de argumentos muy poderosos nos fuerzan a la conclusión de que el libre albedrío no tiene lugar en el universo. De otro, una serie de argumentos poderosos basados en hechos de nuestra experiencia nos inclina a la conclusión de que tiene que haber algo de libre albedrío, puesto que todos nosotros lo experimentamos continuamente.

Hay una solución estándar a este enigma filosófico. De acuerdo con esta solución, el libre albedrío y el de-

terminismo son compatibles entre sí. Desde luego, todo en el mundo está determinado, pero algunas acciones humanas son, sin embargo, libres. Decir que son libres no es negar que estén determinadas; es decir solamente que no están constreñidas. No estamos forzados a hacerlas. Así, por ejemplo, si una persona es forzada a hacer algo a punta de pistola, o sufre alguna compulsión psicológica, entonces su conducta es genuinamente no libre. Pero si por otra parte actúa libremente, si actúa, por así decirlo, de acuerdo con su libre arbitrio, entonces su conducta es libre. Desde luego está también completamente determinada, puesto que todo aspecto de su conducta está determinado por las fuerzas físicas que operan sobre las partículas de que está compuesto su cuerpo, como operan sobre todos los demás cuerpos del universo. Así, la conducta libre existe, pero es solamente una esquinita del mundo determinado; es aquella esquina de la conducta humana determinada de donde ciertos géneros de fuerza y compulsión están ausentes.

Ahora bien, puesto que este punto de vista asevera la compatibilidad del libre albedrío y el determinismo, es usual denominarlo simplemente 'compatibilismo'. Pienso que es inadecuado como solución al problema, y he aquí por qué. El problema sobre el libre albedrío no es sobre si hay o no razones psicológicas internas que causan que hagamos cosas, así como también causas físicas externas y compulsiones internas. Más bien es sobre si las causas de nuestra conducta, cualesquiera que ellas sean, son suficientes o no para *determinar* la conducta de modo que las cosas *tengan que* suceder del modo que suceden.

Hay otra manera de plantear este problema. ¿Es siempre verdadero decir de una persona que podría haber actuado de otra manera, permaneciendo idénticas las otras condiciones? Por ejemplo, dado que una persona eligió votar por los conservadores, ¿podría haber elegido votar por alguno de los otros partidos, permaneciendo idénticas todas las demás condiciones? Ahora bien, el compatibilismo no responde realmente a esta pregunta de un modo que permita alguna oportunidad para la noción ordinaria de libre albedrío. Lo que dice es que toda la conducta está determinada de tal manera que no podría

haber ocurrido de otra manera, permaneciendo idénticas todas las demás condiciones. Todo lo que ha sucedido estaba, de hecho, determinado. Sucede solamente que algunas cosas estaban determinadas por causas psicológicas internas (aquéllas que llamamos nuestras 'razones para actuar') y no por fuerzas externas o compulsiones psicológicas. Así pues, quedamos aún con un problema. ¿Es siempre verdadero decir de un ser humano que él podría haber actuado de otra manera?

El problema que plantea el compatibilismo es, entonces, que no responde a la pregunta: ¿podríamos haber actuado de otra manera, permaneciendo idénticas todas las demás condiciones?, de una manera que sea coherente con nuestra creencia en nuestro propio libre albedrío. El compatibilismo niega, para decirlo brevemente, la sustancia del libre albedrío, mientras que mantiene su caparazón verbal.

Intentemos, entonces, tomar un punto de partida nuevo. He dicho que tenemos una convicción de nuestro propio libre albedrío basada sobre hechos de la experiencia humana. ¿Pero hasta qué punto son fiables esas experiencias? Como mencioné antes, el caso típico, descrito a menudo por los filósofos, que nos inclina a creer en nuestro propio libre albedrío, es un caso en el que nos enfrentamos a un puñado de opciones, razonamos sobre qué es lo mejor que podemos hacer, nos hacemos nuestra composición de lugar y a continuación hacemos lo que hemos decidido hacer.

Pero quizá nuestra creencia de que tales experiencias apoyan la doctrina de la libertad humana es ilusoria. Considérese este tipo de ejemplo. Un experimento de hipnosis típico tiene la forma siguiente. Bajo los efectos de la hipnosis se le hace al paciente una sugerencia poshipnótica. Se le dice, por ejemplo, que haga alguna cosa completamente trivial, inocua, tal como, pongamos por caso, arrastrarse por el suelo. Después que el paciente sale del estado hipnótico, él podría estar conversando, sentado tomando café, cuando de pronto dice algo parecido a '¡Qué suelo más fascinante hay en esta habitación!', o 'Quiero darle un vistazo a esa alfombra', o 'estoy pensando en invertir en revestimientos de suelos y

me gustaría examinar este suelo'. Y a continuación se pone a arrastrarse por el suelo. Ahora bien, el interés de estos casos es que el paciente siempre da alguna razón más o menos adecuada para hacer lo que hace. Esto es, le parece a él mismo que se está comportando libremente. Nosotros, por otra parte, tenemos buenas razones para creer que esta conducta no es libre, en absoluto, que las razones que da para su aparente decisión de arrastrarse por el suelo son irrelevantes, que su conducta estaba determinada con anticipación, que de hecho está en las garras de una sugestión poshipnótica. Cualquiera que conociese los hechos sobre él podría predecir su conducta con anticipación. Ahora bien, una manera de plantear el problema del determinismo, o al menos un aspecto del problema del determinismo es: '¿Es toda la conducta humana semejante a ésta?' ¿Es toda la conducta humana semejante a la de la persona que opera bajo una sugestión poshipnótica?

Pero si tomamos ahora el ejemplo seriamente, parece como si resultase ser un argumento a favor del libre albedrío y no en contra de él. El agente pensaba que estaba actuando libremente, aunque de hecho su conducta estaba determinada, pero parece empíricamente muy poco probable que toda la conducta humana sea semejante a ésta. Algunas veces la gente sufre los efectos de la hipnosis, y algunas veces sabemos que están en las garras de impulsos inconscientes que no pueden controlar. Pero están siempre en situaciones semejantes a éstas? ¿Está toda la conducta determinada por tales compulsiones *psicológicas?* Si intentamos considerar al determinismo psicológico como una afirmación fáctica sobre nuestra conducta, entonces parece ser completamente falsa. La tesis del determinismo psicológico afirma que causas psicológicas anteriores determinan toda nuestra conducta, del modo en que determinan la conducta del sujeto bajo los efectos de la hipnosis, o la del heroinómano. Según este punto de vista, toda la conducta, de una manera u otra, es psicológicamente compulsiva. Pero la evidencia disponible sugiere que tal tesis es falsa. En efecto, nosotros actuamos normalmente sobre la base de nuestros estados intencionales —nuestras creen-

cias, esperanzas, temores, deseos, etc.—, y en este sentido nuestros estados mentales funcionan causalmente. Pero esta forma de causa y efecto no es determinista. Podríamos haber tenido exactamente esos estados mentales y con todo no haber hecho lo que hicimos. Por lo que respecta a las causas psicológicas, podríamos haber actuado de otra manera. Los ejemplos de hipnosis y de conducta psicológicamente compulsiva son en general, por otra parte, patológicas y se distinguen bien de la acción normal libre. Así, psicológicamente hablando, hay oportunidad para la libertad humana.

¿Pero es esta solución un avance sobre el compatibilismo? ¿No estamos diciendo otra vez que sí, que toda conducta está determinada, pero que lo que llamamos conducta libre es el tipo determinado por procesos de pensamientos racionales? Algunas veces los procesos de pensamiento racionales, conscientes, no tienen ninguna influencia, como en el caso de la hipnosis, y algunas veces la tienen, como en los casos normales. Los casos normales son aquéllos en los que decimos que el agente es realmente libre. Pero desde luego esos procesos de pensamiento racionales, normales, están tan determinados como cualquier otra cosa. Así, una vez más, ¿no tenemos el resultado de que todo lo que hacemos estaba enteramente escrito en el libro de la historia billones de años antes de que naciéramos y que, por lo tanto, nada de lo que hacemos es libre en ningún sentido filosófico interesante? Si elegimos llamar libre a nuestra conducta, esto es solamente una cuestión de adoptar una terminología tradicional. Lo mismo que continuamos hablando de 'puestas de sol' incluso aunque sepamos que el sol literalmente no se pone, así continuamos hablando de 'actuar de acuerdo con nuestro libre albedrío', aunque no haya tal fenómeno.

Una manera de examinar una tesis filosófica, o cualquier otro género de tesis por lo que respecta a esto es preguntar '¿cuál es la importancia de esto? ¿Hasta qué punto el mundo sería diferente si esta tesis fuera verdadera, cómo opuesto a como sería el mundo si esta tesis fuese falsa?' Parte del atractivo del determinismo, creo, es que parece consecuente con el modo en que el mundo

procede de hecho, al menos en la medida en que conocemos algo sobre él por la física. Esto es, si el determinismo fuese verdadero, entonces el mundo procedería más o menos de la misma manera que procede, con la única diferencia de que ciertas creencias nuestras acerca de cómo procede serían falsas. Esas creencias son importantes para nosotros porque tienen que ver con la creencia de que podríamos haber hecho las cosas de manera diferente a como de hecho las hicimos. Y esta creencia se conecta a su vez con creencias sobre responsabilidad moral y sobde nuestra propia naturaleza como personas. Pero si el libertarismo, que es la tesis del libre albedrío, fuese verdadero, parece que tendríamos que hacer algunos cambios radicales acerca de nuestras creencias sobre el mundo. Para que tengamos libertad radical, parece como si tuviéramos que postular que dentro de cada uno de nosotros había ya un yo capaz de intervenir en el orden causal de la naturaleza. Esto es, parece como si tuviéramos que contener alguna entidad que fuese capaz de hacer que las moléculas se desviasen de sus trayectorias. No sé si tal punto de vista es ni siquiera inteligible, pero no es ciertamente consistente con lo que sabemos sobre cómo el mundo funciona de acuerdo con la física. Y no hay la menor prueba de que debamos abandonar la teoría física en favor de tal punto de vista.

Hasta ahora parece que no estamos yendo exactamente a ninguna parte en nuestro esfuerzo por resolver el conflicto entre el determinismo y la creencia en el libre albedrío. La ciencia no deja lugar para el libre albedrío, y el indeterminismo en física no le ofrece ningún apoyo. Por otra parte, no somos capaces de abandonar la creencia en el libre albedrío. Investiguemos estos puntos un poco más.

¿Por qué exactamente no hay lugar para el libre albedrío según el punto de vista científico contemporáneo? Nuestros mecanismos explicativos básicos en física funcionan de abajo arriba. Es decir, explicamos la conducta de características superficiales de un fenómeno tales como la transparencia del vidrio o la liquidez del agua, en términos de la conducta de micropartículas tales como las moléculas. Y la relación de la mente con el ce-

rebro es un ejemplo de tal relación. Los rasgos mentales están causados por, y realizados en, los fenómenos neurofisiológicos, como he discutido en el primer capítulo. Pero tenemos causación de la mente al cuerpo; esto es, tenemos causación de arriba abajo a través de una porción de tiempo, y tenemos causación de arriba abajo a través del tiempo porque el nivel superior y el nivel inferior van juntos. Así, por ejemplo, supongamos que quiero causar la descarga del neurotransmisor acetilcolina en las placas terminales del axón de mis neuronas motoras. Puedo hacerlo decidiendo simplemente levantar mi brazo, y levantándolo a continuación. Aquí el evento mental, la intención de levantar el brazo, causa el evento físico, la descarga de acetilcolina, un caso de causación de arriba abajo si alguna vez hubo alguno. Pero la causación de arriba abajo funciona solamente porque los eventos mentales, para empezar, están fundados en la neurofisiología. Así, correspondiendo a la descripción de las relaciones causales que van de arriba abajo, hay otra descripción de la misma serie de eventos donde las relaciones causales rebotan enteramente en la parte de abajo. Esto es, son completamente un asunto de neuronas y de activación de neuronas en las sinapsis, etcétera. En la medida en que aceptamos esta concepción de cómo funciona la naturaleza, entonces no parece que haya ninguna oportunidad para el libre albedrío, puesto que según esta concepción la mente puede afectar solamente a la naturaleza en tanto que es parte de la naturaleza, sus rasgos están determinados a los microniveles básicos de la física.

Este es un punto absolutamente fundamental de este capítulo, de modo que lo voy a repetir. La forma de determinismo que es, en última instancia, preocupante, no es el determinismo psicológico. La idea de que nuestros estados mentales son suficientes para determinar todo lo que hacemos es probablemente sólo falsa. La forma de determinismo preocupante es más básica y fundamental. Puesto que todos los rasgos superficiales del mundo están causados enteramente por y realizados en sistemas de microelementos, la conducta de los microelementos es suficiente para determinar todo lo que su-

cede. Tal cuadro del mundo 'de abajo arriba' permite causación de arriba abajo (nuestras mentes, por ejemplo, pueden afectar a nuestros cuerpos). Pero la causación de arriba abajo solamente funciona porque el nivel superior está ya causado por, y realizado en, los niveles inferiores.

Bien, volvamos entonces a la siguiente pregunta obvia. ¿Qué es lo que pasa en nuestra experiencia que nos hace imposible abandonar la creencia en el libre albedrío? Si la libertad es una ilusión, ¿por qué es una ilusión que parece que somos incapaces de abandonar? La primera cosa que tenemos que observar sobre nuestra concepción de la libertad humana es que está ligada esencialmente a la conciencia. Sólo atribuimos libertad a los seres conscientes. Si, por ejemplo, alguien construyese un robot que creyésemos que fuera totalmente inconsciente, jamás sentiríamos ninguna inclinación de decir de él que era libre. Incluso si encontráramos que su conducta era aleatoria e impredecible, no diríamos que estaba actuando libremente en el sentido en que pensamos que nosotros mismos actuamos libremente. Si, por otra parte, alguien construyese un robot tal que nos convenciésemos de que tenía conciencia, en el mismo sentido que nosotros la tenemos, entonces sería cuando menos una cuestión abierta si el robot tenía o no libre albedrío.

El segundo punto que hemos de observar consiste en que no es justamente ningún estado de conciencia el que nos da la convicción de la libertad humana. Si la vida consistiese enteramente en la recepción de percepciones pasivas, entonces me parece que ni siquiera nos formaríamos nunca la idea de libertad humana. Si nos imaginamos a nosotros mismos totalmente inmóviles, incapaces incluso de determinar el curso de nuestros propios pensamientos, pero recibiendo con todo estímulos, por ejemplo, sensaciones periódicas levemente dolorosas, no tendríamos la más ligera inclinación a concluir que teníamos libre albedrío.

Dije anteriormente que la mayor parte de los filósofos piensan que la convicción de la libertad humana está ligada esencialmente con el proceso racional de tomar decisiones. Pero pienso que esto es sólo parcialmente ver-

dadero. De hecho, sopesar razones es solamente un caso muy especial de la experiencia que nos da la convicción de libertad. La experiencia característica que nos da la convicción de la libertad humana, y ésta es una experiencia de la cual somos incapaces de apartar la convicción de libertad, es la experiencia de ocuparnos en acciones humanas, intencionales, voluntarias. En nuestra discusión de la intencionalidad nos hemos concentrado en aquella forma de intencionalidad que consiste en intenciones en la acción conscientes, intencionalidad que es causal del modo en que la he descrito, y cuyas condiciones de satisfacción son que ocurran ciertos movimientos corporales, y que ocurran en tanto que causados por esa misma intención en la acción. Es esta experiencia la que constituye los cimientos de nuestra creencia en el libre albedrío. ¿Por qué? Reflexionemos muy cuidadosamente en el carácter de las experiencias que se tienen cuando uno está ocupado en las acciones humanas ordinarias, de todos los días, normales. Se experimentará, empotrada dentro de esas experiencias, la posibilidad de desarrollos de acción alternativos. Si uno levanta su brazo o atraviesa la habitación para tomar un trago de agua, verá que en cualquier punto de la experiencia se tiene una sensación de que están abiertos desarrollos alternativos de la acción.

Si se intentase explicar en palabras, la diferencia entre la experiencia de percibir y la experiencia de actuar consiste en que al percibir se tiene el sentido: 'Esto me está sucediendo a mí', y al actuar se tiene el sentido: 'Estoy haciendo que esto suceda.' Pero el sentido de 'Estoy haciendo que esto suceda' comporta el sentido de que 'Yo podría estar haciendo algo distinto'. En la conducta normal, cada cosa que hacemos comporta la convicción, válida o inválida, de que podríamos estar haciendo algo distinto perfectamente aquí y ahora; esto es, permaneciendo idénticas todas las demás condiciones. Esto, propongo yo, es la fuente de nuestra inquebrantable convicción de nuestro propio libre albedrío. Es quizá importante subrayar que estoy discutiendo la acción humana normal. Si uno está en las garras de una gran pasión, si uno está preso de furor, por ejemplo, uno pierde este sentido de libertad

y puede incluso de sorprenderse de descubrir lo que está haciendo.

Una vez que observamos este rasgo de la experiencia de actuar, un gran número de los problemáticos fenómenos que he mencionado anteriormente se explican fácilmente. ¿Por qué, por ejemplo, tenemos la sensación de que la persona en el caso de la sugestión poshipnótica no está actuando libremente en el sentido en que nosotros lo estamos haciendo, incluso aunque él pudiera pensar que está actuando libremente? La razón es que en un sentido importante él no sabe lo que está haciendo. Su intención-en-la-acción efectiva es totalmente inconsciente. Las opciones que ve como disponibles para él son irrelevantes para la motivación efectiva de su acción. Obsérvese también que los ejemplos compatibilistas de conducta 'forzada' incluyen, sin embargo, en muchos casos, la experiencia de libertad. Si alguien me dice que haga algo a punta de pistola, incluso en tal caso tengo una experiencia que tiene empotrados dentro de ella desarrollos alternativos de la acción. Si, por ejemplo, se me ordena atravesar la habitación a punta de pistola, con todo es parte de mi experiencia el sentir que en cualquier paso tengo abierta la posibilidad de hacer algo distinto. La experiencia de libertad es entonces un componente esencial de cualquier caso en el que se actúa con una intención.

De nuevo, esto puede verse si se contrasta el caso normal de una acción con los casos de Penfield, donde la estimulación del córtex motor produce un movimiento involuntario del brazo o la pierna. En tal caso, el paciente experimenta al movimiento pasivamente, como si tuviera la experiencia de un sonido o una sensación de dolor. A diferencia de las acciones intencionales, no hay opciones empotradas en la experiencia. Para ver este punto claramente, intente usted imaginar que una porción de su vida fue algo parecido a los experimentos de Penfield a gran escala. En lugar de atravesar la habitación, usted sentía simplemente que su cuerpo se movía a través de la habitación; en lugar de hablar, usted simplemente oía y sentía que las palabras salían de su boca. Imagínese que sus experiencias son las de una marioneta

puramente pasiva, pero consciente, y usted se habrá imaginado sin la experiencia de libertad. Pero en el caso típico de acción intencional, no hay modo alguno en que podamos separar la experiencia de libertad. Es una parte esencial de la experiencia de actuar.

Esto explica, creo, por qué no podemos abandonar nuestra convicción de libertad. Encontramos que es fácil abandonar la convicción de que la tierra es plana tan pronto como entendemos la evidencia para la teoría heliocéntrica del sistema solar. De modo parecido cuando miramos una puesta de sol no nos sentimos obligados a creer que, a pesar de las apariencias. el sol se está poniendo detrás de la tierra, creemos que la apariencia de que el sol se está poniendo es simplemente una ilusión creada por la rotación de la tierra. En cada caso es posible abandonar una convicción de sentido común, puesto que la hipótesis que la reemplaza da cuenta de que las experiencias que, en primer lugar, llevaron a esa convicción, así como explica una gran cantidad de otros hechos de los que no es capaz de dar cuenta el punto de vista del sentido común. Esta es la razón por la que abandonamos la creencia en una tierra plana y en unas 'puestas de sol' literales a favor de la concepción copernicana del sistema solar. Pero igualmente no podemos abandonar la convicción de libertad, puesto que esa convicción está empotrada en toda acción intencional consciente, normal. Y usamos esta convicción al identificar y explicar acciones. Este sentido de libertad no es solamente un rasgo de deliberación, sino que es parte de cualquier acción, ya sea premeditada o espontánea. El núcleo de esto no tiene esencialmente nada que ver con la deliberación: la deliberación es simplemente un caso especial.

No navegamos bajo la suposición de que la tierra es plana, incluso aunque la tierra parece plana, pero actuamos bajo la suposición de libertad. De hecho no podemos actuar de otra manera que bajo la suposición de libertad, sin importarnos lo mucho que aprendamos sobre cómo funciona el mundo como un sistema físico determinista.

Podemos ahora extraer las conclusiones que están implícitas en esta discusión. En primer lugar, si la preocupación sobre el determinismo es la preocupación de

que toda nuestra conducta es psicológicamente compulsiva, entonces parece que la preocupación está injustificada. En la medida en que el determinismo psicológico es un hipótesis empírica igual que cualquier otra, la evidencia de que en la actualidad disponemos sugiere que es falsa. Así, esto nos da una forma modificada de compatibilidad. Nos proporciona el punto de vista de que el libertarismo psicológico es compatible con el determinismo físico.

En segundo lugar, nos da incluso un sentido de 'podría haber' en el que la conducta de las personas, aunque determinada, es tal que, en este sentido, podrían haber actuado de otra manera. El sentido es simplemente que por lo que respecta a los factores *psicológicos* podrían haber actuado de otra manera. Las nociones de capacidad, de lo que somos capaces de hacer y de lo que podríamos haber hecho, son relativas a menudo a algún conjunto tal de criterios. Por ejemplo, yo podría haber votado por Carter en las elecciones americanas de 1980, incluso si no lo hice; pero no podría haber votado por George Washington. No era candidato. Así, hay un sentido de 'podría haber', en el que tenía disponibles una serie de elecciones, y en ese sentido había una gran cantidad de cosas que yo podría haber hecho, permaneciendo idénticas todas las demás cosas, y que no hice. Similarmente, puesto que los factores psicológicos que operan sobre mí no me compelen nunca, o incluso en general, a comportarnos de una manera determinada, a menudo, psicológicamente hablando, podría haber hecho algo diferente de lo que de hecho hice.

Pero en tercer lugar, esta forma de compatibilismo no nos da, con todo, nada parecido a la resolución del conflicto entre libertad y determinismo que nuestro impulso hacia el libertarismo radical exige realmente. En la medida en que aceptamos la concepción de abajo-arriba de la explicación física, y esta es una concepción en la que están basados los pasados trescientos años de ciencia, los hechos psicológicos sobre nosotros mismos, al igual que cualesquiera otros hechos de nivel superior, son explicables causalmente de manera completa en términos de, y están enteramente realizados en, sistemas de

elementos al nivel microfísico fundamental. Nuestra concepción de la realidad física no deja lugar simplemente para la libertad radical.

En cuarto lugar, y finalmente, por razones que yo realmente no entiendo, la evolución nos ha dado una forma de experiencia de acción voluntaria donde la experiencia de libertad, es decir, la experiencia del sentido de posibilidades alternativas, está empotrada en la misma estructura de la conducta humana intencional, voluntaria, consciente. Por esta razón, creo que ni esta discusión ni ninguna otra nos convencerá jamás de que nuestra conducta no es libre.

Mi aspiración en este libro ha sido intentar caracterizar las relaciones entre la concepción que tenemos de nosotros mismos como agentes racionales, libres, conscientes, cuidadosos, y la concepción que tenemos del mundo como algo que consta de partículas físicas sin mente, carentes de significado. Resulta tentador pensar que justamente en la medida en que hemos descubierto que extensas porciones de sentido común no representan adecuadamente cómo funciona realmente el mundo, en esa misma medida podríamos descubrir que nuestra concepción de nosotros mismos y de nuestra conducta es enteramente falsa. Pero hay límites a esta posibilidad. La distinción entre realidad y apariencia no se puede aplicar a la misma existencia de la conciencia. Pues si me parece que soy consciente, entonces *soy* consciente. Podríamos descubrir todo género de cosas sorprendentes sobre nosotros mismos y sobre nuestra conducta, pero no podemos descubrir que no tenemos mentes, que éstas no contienen estados mentales intencionales, subjetivos, conscientes, ni podríamos descubrir que, al menos, intentamos ocuparnos en acciones intencionales, libres, voluntarias. El problema que me he planteado no es probar la existencia de esas cosas, sino examinar su estatus y sus implicaciones para nuestras concepciones del resto de la naturaleza. Mi tema general ha sido que, con ciertas excepciones importantes, nuestra concepción mentalista común de nosotros mismos es perfectamente consistente con nuestra concepción de la naturaleza como un sistema físico.

MIENTRAS que el sentido común nos presenta como seres racionales, conscientes y libres, la ciencia nos informa de que el Universo físico en que operamos, y que constituimos, es un agregado de partículas inconscientes.

¿Qué sabemos del punto de unión entre nuestra mente y el mundo físico —el cerebro—? Muy poco. Sólo recientemente se abre paso una atrevida analogía que relaciona el cerebro con un ordenador digital. Ciertos científicos postulan que las máquinas pueden —o podrán— pensar, y la ciencia de estos mecanismos ya tiene un nombre: Inteligencia Artificial. ¿Resuelve la I. A. los problemas planteados por la relación mente/cerebro?

John Searle, profesor de Filosofía en la Universidad de Berkely, mantiene en la actualidad la postura más brillante en contra de los abusos de la Inteligencia Artificial. En su actividad como filósofo del lenguaje, Searle adoptó el punto de vista de la pragmática —esencialmente anti-chomskyano— con sus *Actos de habla* (en esta misma colección).